# Understanding compound-induced histopathology in rat liver using gene expression network methods

**Benjamin Alexander-Dann**

King's College

University of Cambridge

Submitted September 2019

UNIVERSITY OF CAMBRIDGE

This dissertation is submitted for the degree of Doctor of Philosophy

# Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the Degree Committee for the Faculty for Physics and Chemistry

# Understanding compound-induced histopathology in rat liver using gene expression network methods

## Summary

Current drug discovery is a lengthy and costly pipeline; it takes between twelve and fifteen years and costs $1-2 billion (USD). As such, any compound failures represent a sunk cost – exacerbated if such failures occur later in the pipeline. Compound and drug induced liver injury is a significant cause of failures. Current progress to tackle this is based on systems biology, and so falls within the field of toxicogenomics. As such, the databases in the public domain are crucial to progress. DrugMatrix and Open TG-GATEs were identified as containing large, *in vivo* transcriptomics data with histopathological observations as endpoints, a proxy for toxicity. Due to the size and chemical variety of the databases, data-driven methods led to the novel creation of histopathology signatures, which accounts for dependence between histopathology observations (Chapter 2).

Six toxic groups were determined for DrugMatrix and 13 for Open TG-GATEs, and were analysed with a view to enable classification, namely, what is revealed in the gene expression profiles when the histopathology phenotype was present. This led to determining gene-phenotype associations, both known and novel. An example of a novel association was the match of the histopathology signature of 'glycogen accumulation, mixed infiltration and lymphocytic inflammatory cell infiltration' to fructose metabolism, gluconeogenesis, and chemokine response pathways (Chapter 3). Concordance was found between histopathologically related toxicity groups between databases and their co-expression networks.

From here, the co-expression network methods were applied to determine the concordance of gene expression across time (one day to four/five days), database (DrugMatrix and Open TG-GATEs), and toxicity group. This found underlying biological terms such as RNA transport, ribosome biogenesis and translation as well as toxicity-specific terms (aminoacyl t-RNA synthesis in metabolic processes for the histopathological observations of glycogen accumulation, cellular infiltrate, hepatocellular necrosis and fatty change in liver. Crucially, this work determined that the toxic group membership plays a more significant role in gene co-expression networks compared to the time point of the gene expression measurement (Chapter 4).

In conclusion, data driven clustering was performed to create histopathology signatures. Using these, the usefulness of transcriptomics data was determined both to classify toxic state

(gene expression data measured when the phenotype was present) and to determine how consistent it is over time scales. This work provided a framework for the comparison of co-expression networks for the deconvolution of gene expression data with respect to a phenotype.

Benjamin Alexander-Dann

# Acknowledgements

I would like to acknowledge Dr. Andreas Bender for giving me the opportunity to pursue this PhD in his group, and Dr. Tim James, my collaborator, for his support and help. In addition, all the members of the Bender group (past and current) have created a fantastic and supportive group. In particular, I would like to thank Dr. George Drakakis, Dr. Krishna Bulusu, Dr. Dezso Modos, Dr. Lewis Mervin, Dr. Fatima Baldo, Dr. Erin Oerton, Dr. Christoph Schlaffner, Dr. Stephanie Ashenden, Dr Kathryn Giblin and Mr Chad Allen.

On a personal note, I owe everlasting thanks to my family, to my father Tim, my brothers Mark and James, my sister Pia and my grandmother Jill. This would not have been possible without you all.

# Publications

The following publications have been generated during this PhD:

**Alexander-Dann B**, Pruteanu LL, Oerton E, et al. Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. Mol Omics. 2018;14(4):218–236. doi:10.1039/c8mo00042e
[This paper forms the base of the introduction]

Other publications not included in this thesis:

Drakakis, G., Cortés-Ciriano, I., **Alexander-Dann, B**, & Bender, A. Elucidating compound mechanism of action and predicting cytotoxicity using machine learning approaches, taking prediction confidence into account. Current Protocols in Chemical Biology, 2019, 11, e73. doi: 10.1002/cpch.73

# 1. Introduction to Toxicity within Current Drug Discovery

Current drug discovery is a lengthy and costly pipeline; it takes between twelve and fifteen years and costs $1-2 billion (USD).[1–3] As such, any compound failures represent a sunk cost – exacerbated if such failures occur later in the pipeline. There has been a significant increase in the investment and technologies for drug development but these have failed to increase the rate of approved drugs. This is the so called "innovation gap" and has two main drivers: efficacy and safety.[4,5] Whilst there are numerous reasons for the former, this work focuses specifically on the latter.

The aim of this work is to identify biological indicators of unsafe compounds, using chemical structure, gene expression and biological pathways. This is two fold: to identify how toxic end points are reached (biological understanding) and to determine whether particular biomarkers can identified for effective compound screening in drug development (predictive toxicology).

The toxicity of drugs is a significantly regulated area, with changes enforced regularly, particularly after high profile cases, such as the thalidomide scandal.[6] This scandal, termed the "biggest medical disaster in history" lead to severe birth defects for 10,000 children.[7] This was not an isolated incident. 462 medicinal products were removed from the marketplace between 1953 and 2013, with the most common reason being hepatotoxicity.[8] In addition to these withdrawals, there were postmarketing safety events for 32% of the 222 novel therapeutics that were approved by the United States of America's Food and Drug Administration (FDA) between 2001-2010. Furthermore, safety issues accounted for 24% of clinical trial failures between 2013 to 2015 – the second highest cause, behind efficacy.[9] The issue is mirrored in a non-clinical setting: 40% of compound failures in preclinical studies were due to toxicity.[10] From these failures (from post marketing to preclinical studies), the importance of safety and toxicology studies is shown to be paramount. There are four main types of toxicity: geno-, nephron-, cardio-, and hepato- toxicity. These are reviewed briefly in Table 1-1.

Compound- and drug-induced liver injury, a part of hepatotoxicity is of particular relevance for drug development, due to the liver's function in the metabolism of xenobiotic substances.

*Table 1-1 The four main types of unacceptable toxicity failure: genotoxicity, nephrotoxicity, cardiotoxicity and hepatotoxicity.*

| Type | Definition |
|---|---|
| Genotoxicity | Destructive effect on a cell's genetic material (DNA/RNA). These toxins are split into three main groups: carcinogens (cancer-causing), mutations (causing mutations) or teratogens (birth defect causing).[11] |
| Nephrotoxicity | Damage or destruction of kidney function causing kidney-specific detoxification and excretion failures. These toxins are split into mechanism based groups: changes in glomerular hemodynamics, tubular cell toxicity, inflammation, crystal nephropathy, rhabdomyolysis, and thrombotic microangiopathy.[12] |
| Cardiotoxicity | Damage to heart muscle or heart electrophysiology dysfunction. These structural and functional toxicities are difficult to separate as structural damage often results in reduced function and *vice versa.*[13] |
| Hepatotoxicity | Injury to hepatocytes (liver cells) or bile duct cells. Main mechanisms include bile acid-induced hepatocyte apoptosis, adhesion molecules and oxidant stress in inflammatory liver injury, cytochrome P4502E1-dependent toxicity, peroxynitrite (its formation correlates with necrosis), and mitochondrial dysfunction.[14] |

## 1.1 Compound and Drug Induced Liver Injury

Compound and drug induced liver injury (DILI) in human patients is a diagnosis of exclusion, within human patients (i.e. all other plausible causes have been ruled out).[15] It manifests clinically as cholestasis and/or hepatocellular damage, which can include changes

to metabolism, transport protein function and direct hepatocellular damage.[16,17] Whilst it has significant human implications, pre-human trials are the focus of this work. It is proposed that the pathogenesis of DILI is split into three main mechanisms: direct cell stress, direct mitochondrial stress, and immune reactions (extrinsic to cell).[18] The mechanisms and phenotypes of DILI are discussed in the following sections.

### 1.1.1. Mechanisms of drug-induced liver injury

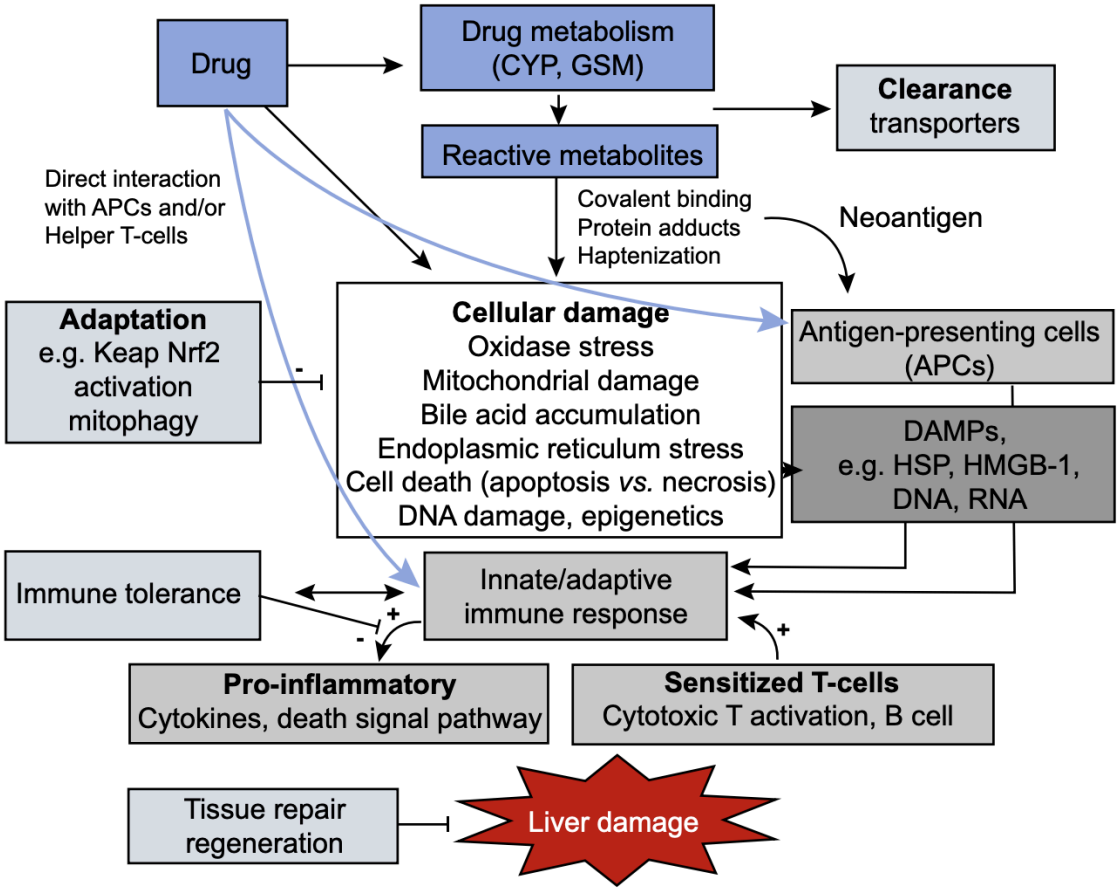



*Figure 1-1 Mechanisms of drug-induced liver injury. Drugs either directly cause cytotoxicity or via immune response with antigen-presenting cells. The drugs, or reactive metabolites, can cause damage through oxidase stress, mitochondrial damage, bile acid accumulation, endoplasmic reticulum stress, cell death or DNA damage. Affected cells can produce danger signals, such as "danger associated molecular patterns" (DAMPs), which favour the release of cytokines. This figure was adapted from Chen et al.*[19]

A toxic drug (or compound) either directly causes cellular damage or indirectly, via activation of the immune system, as shown in Figure 1-1. The drug, or its reactive metabolites formed in the liver, cause either oxidase stress, mitochondrial damage, bile acid metabolism, endoplasmic reticulum stress, cell death or DNA damage. Antigen-presenting

cells (a heterogeneous group of cells that mediate immune response by processing and presenting antigens for lymphocytes) cause injury to hepatocytes, which in turn release so-called “danger signals” (e.g. damage associated molecular patterns molecules, DAMPs).

Oxidase stress occurs due to an imbalance of reactive oxygen species (ROS) formation (via the c-Jun N-terminal kinase, JNK) and detoxification via Nrf2/Keap1.[20] More specifically Nrf2 plays an important part in liver injury: it both promotes the HO-1 gene which encourages autophagy (the act of cleaning out damaged cells) and so protects the liver, and it regulates antioxidant-relevant genes, However, Keap1 can bind to Nrf2 and inhibit its activity, resulting in liver injury via the Trx-PI3K/AKT-HIF1-HO-1/CyclinD1 signalling pathway. This is shown in Figure 1-2. Increased ROS has been linked to direct damage to DNA, lipids, proteins, enzymes, and intracellular glutathione depletion.[21]

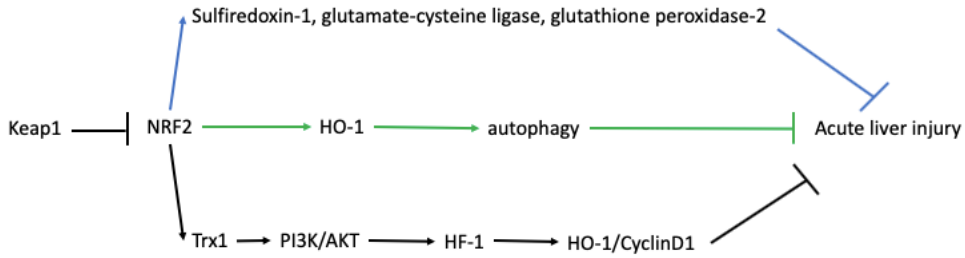



*Figure 1-2 The role of Nrf2 in acute liver injury. Nrf2 protects the liver through regulating antioxidant genes including sulfiredoxin-1, glutamate-cysteine ligase, and glutathione peroxidase-2 (blue lines), and promoting the HO-1 gene and hence autophagy. However, Keap1 can bind to Nrf2 and inhibit it and the Trx-PI3K/AKT-HIF1-HO-1/CyclinD1 signalling pathway and so promoting liver injury. This figure has been adapted from Xu et al.*[22]

Mitochondrial damage plays a critical role in both steatosis and hepatic necrosis via activation of cellular death pathways. Crucially, cell death here is an active process, not a passive biochemical overwhelming of cells.[23] Bile acid accumulation occurs when bile acid (the production of which is one of the primary functions of the liver) is not adequately transported away from the liver. The inhibition of bile salt export pump, am ATP-dependant transporter, has been implicated in the accumulation of bile salts (which are cytotoxic), which in turn activate oxidative stress and FAS pathways.[24] These proposed mechanisms are closely related to the histopathology observations associated with liver damage).

### 1.1.2 Histopathology observations to characterise drug-induced liver injury

Histopathology is often used to characterise DILI (and, indeed, all forms of disease states). The microscopic observations can even be used in an unsupervised manner to define disease states.[25] It is defined as "the microscopic study of animal and plant tissues by the staining, sectioning and examining under microscope".[26] These observations are expert based but considerable work there has been done by the eTox Consortium to create the 'histopathology ontology' (HPATH) which standardises terms and their relation to each other.[27] The characteristics for DILI and liver injury more generally are summarised in Table 1-2. These phenotypes are used to define end points in toxicity studies and so their definition is crucial for comparisons between different experiments. Additionally, these terms are used in this work to define toxic groups.

*Table 1-2: Histopathology observations that are common in liver injury. The observations are taken from Robert Maronpot[28] and the descriptions from the Histopathology Ontology which was developed by the eTox Consortium[27]*

| Histopathology Observation | Description |
|---|---|
| Glycogen deposition | Irregular and poorly defined clear spaces in the cytoplasm (rarefaction) usually with centrally located nuclei. |
| Fatty change | Hepatocellular vacuolation is usually lipidic in nature, nevertheless, vacuoles may develop from increased intracellular fluid contents within vesicles and/or by swelling of cytoplasmic organelles; Fatty change can also be observed in combination with other hepatotoxic injuries (e.g., chronic liver toxicity, degeneration, inflammation, and necrosis) or nutritional disturbance (e.g. diet, vitamin A excess) in both animals and man. |
| Pigmentation | Pigmentation in most tissues consists of lipofuscin and/or hemosiderin, usually present in interstitial macrophages. Nevertheless other cells are capable of accumulating pigments (e.g. hepatocytes, renal tubular cells). Pigments may also originate from specific compounds or their metabolites. Special stains and/or other investigations are |

| | |
|---|---|
| | necessary to further characterize the origin of the pigments in a given tissue. |
| Degeneration | Degeneration is a deterioration of living cells following an insult, with the possibility to reverse after removal of the insult. Degenerated cells show morphological changes (e.g. cell swelling, increased eosinophila/basophila of the cytoplasm) which are in context of biochemical, metabolic and/or mechanical disruption of the cell physiology and integrity. |
| Cell death | Grouping of all changes affecting individual or group of cells, characterized by partial or total overwhelming of the cell capacity to maintain a biochemical and morphological equilibrium, resulting in various type of morphologies. |
| Hypertrophy | Increase in volume of a tissue or organ produced entirely by enlargement of existing cells, without contribution of generation of new cells (such as in hyperplasia). Most organs undergoing hypertrophy can have an increase in cell number as well, to a certain extent. |
| Karyomegaly | An increase in nuclear size and is occasionally noted in rat tubular epithelium. it is presumed to represent repeated nucleic acid replication without nuclear divisions or cytokinesis but its pathogenesis is uncertain. |
| Inflammatory cell infiltrates | When the various combinations of inflammatory cell infiltrates occur in the absence of other features of inflammation |
| Proliferation response: non-neoplastic (hyperplasia) | Nonneoplastic proliferation of hepatocytes which appear unaltered but may have slightly basophilic cytoplasm and/or prominent nuclei. An evidence of prior or ongoing hepatocellular damage is present. |
| Proliferation response: neoplastic | Neoplastic cells are said to be transformed because they continue to replicate, apparently oblivious to the regulatory influences that control normal cell growth. This is caused by DNA mutations that are (for the most part) acquired |

| | |
|---|---|
| | spontaneously or induced by environmental insults. These genetic but also epigenetic changes alter the expression or function of key genes that regulate fundamental cellular processes, such as growth, survival, and senescence, resulting in an abnormal growth. |

Despite knowledge of some of the mechanisms in liver injury and the associated phenotype changes, liver toxicity remains an issue. A prevailing thought that has led to the current issues in drug development is the "target focused drug design". The majority of drug development focuses on the identification of a protein target that is associated with a particular disease or phenotype.[4] In this approach, a biologically relevant target is identified and modelled. Compounds are created and optimised to hit this target and modulate its activity towards the desired effect. Compounds are then progressed through the pipeline, but human toxicity is assessed later. The case has been made for the inclusion of toxicology studies in the exploratory phase of drug development.[29,30] This relies on inherent target safety, chemical series safety and safety in compound lead optimisation. Inherent target safety is mostly concerned with novel targets. For example, anti-cancer compounds designed to hit proteins on the RAS-RAF-MEK-ERK pathway for anti-tumour potential result in skin toxicity, akin to the toxicity seen with multikinase and EGFR inhibitors.[31] This is almost expected as RAF and MEK are major downstream mediators of EGFR signalling.[32] Chemical series safety and safety lead optimisation relate to whether particular physiochemical properties that are known toxicants are included in the optimisation of target modulation.[33]

However, the sole focus on physiochemical descriptors is not sufficient for improved compound attrition rates.[10] Whilst control of these physiochemical properties is important, they fail to correlate well or meaningfully with the complicated and nuanced toxicity end points.

The case for "systems biology" approaches and predictive models of late stage toxicity is clear. This is used in the field of toxicogenomics, defined as "combin[ing] toxicology with information-dense genomic technologies to integrate toxicant-specific alterations in gene, protein and metabolite expression patterns with phenotypic responses of cells, tissues and organisms".[34] The field of toxicogenomics is discussed in Section 1.3. A key part to toxicogenomics is the use of 'omics data.

## 1.2 Measurement of gene expression data

Gene expression measurement is the measurement of RNA, produced from the transcription of DNA and translated to create polypeptides.[35] There are three main methods to measure this data: real-time quantitative polymerase chain reaction (RT-qPCR), microarray analysis and RNA sequencing (RNA-Seq). These have been recently reviewed here [36]. RT-qPCR is the most sensitive of the three but is a more intensive protocol and can only be performed for a limited number of genes. As such, it tends to be used to validate the expression measured from the other, high throughput methods.

Microarrays were developed in the 1990s.[37] A microarray is a series of wells on a chip, to which probes corresponding to known genomic locations of an organism of interest have been bound, as shown in Figure 1-3. These probes are selectively designed nucleotide sequences. It is to these probes that fluorescently tagged reverse transcribed sample cDNA, made from the sample RNA of interest, bind. The location (i.e. which probe) and intensity of fluorescence determine the type and amount of RNA present in the original sample.[38] Due to the nature of the design, only known genes (i.e. those whose RNA product that binds to specifically designed probes) can be measured, limiting *de novo* studies. Therefore lncRNA and miRNA are not suitable for this method, as they are part of less well known genomic regions. Additionally, it assumes that a perfect match occurs between the probes and cDNA, and that no probe is saturated (i.e. no further binding possible), leading to noise in the measurements.[39] Despite this, microarray technologies produce concordant, biologically relevant signals.[40] It is an established technique and there are numerous computational packages encoding algorithms for the analysis of microarray platforms.

In comparison, RNA-Seq does not measure fluorescence, but directly counts the amount of reverse transcribed cDNA. Therefore, it can detect previously unknown sequences, and so is more valid for determining unknown genomic sequences or novel transcripts. Additionally, it can measure lower levels of cDNA than microarray experiments. However, it is currently a more expensive method (although the cost gap is diminishing).

The concordance between microarrays and RNA-Seq has been studied in the context of liver toxicity in rats.[41] Compound-induced gene expression data from 5 known toxicants was measured using both microarray and RNA-Seq platforms. High levels of agreement were observed (78% overlap of differentially expressed genes) but RNA-Seq did find more differentially expressed genes and enriched pathways.

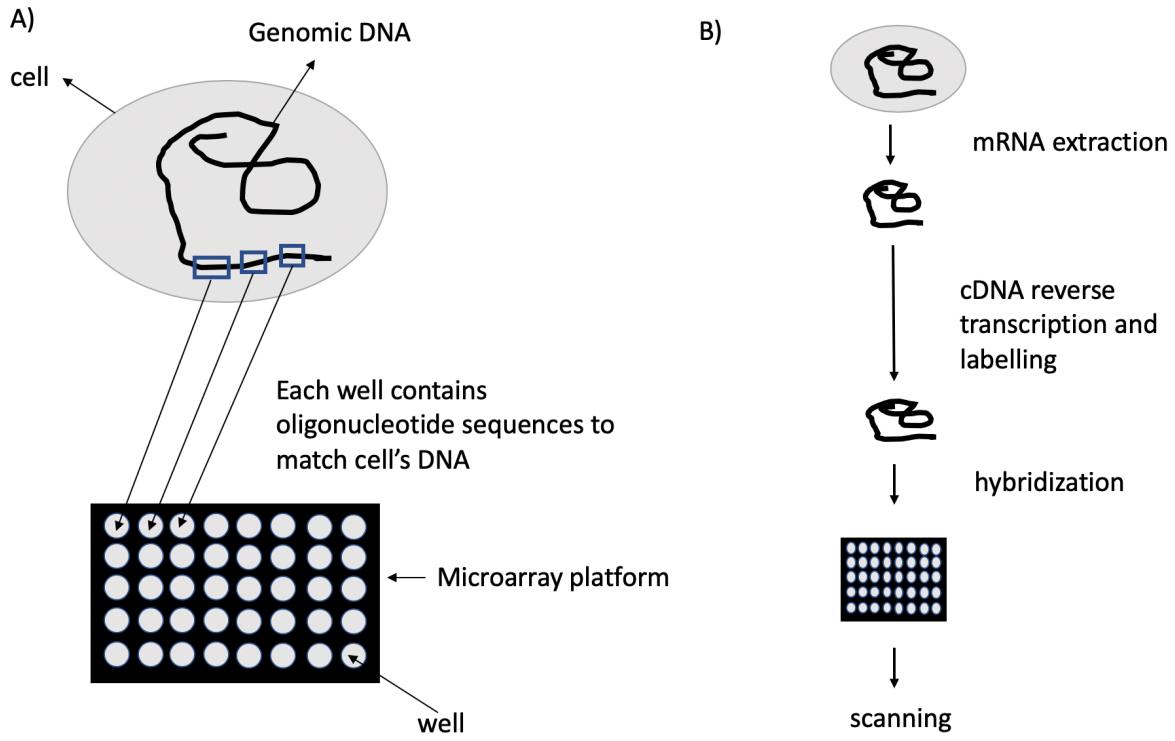



*Figure 1-3 Gene expression measurement using microarray platform technology, A) shows the design of the platform, which is a series of wells that contain specific oligonucleotide sequences to match up with the DNA of chosen sample. B) shows the processes of expression measurement.*

## 1.3 The field of toxicogenomics

The field of toxicogenomics is defined as "combin[ing] toxicology with information-dense genomic technologies to integrate toxicant-specific alterations in gene, protein and metabolite expression patterns with phenotypic responses of cells, tissues and organisms".[34] It has numerous aims, namely to predict toxic endpoints, to determine the mechanism of toxic action, and to identify biomarkers of toxicity.[42] Data limitations have narrowed the scope of the field to transcriptomics data (which are more numerous and systematic).[43]

A prominent, early limitation of this field was the lack of large-scale, public, suitable databases. However, this has changed over the last eight years, with the introduction of two toxicogenomics-specific databases; namely DrugMatrix[44] and Open TG-GATEs.[45] Both of these contain *in vivo*, compound-induced gene expression data, along with histopathology

annotations. Additionally, complementary transcriptomics databases provide compound-induced gene expression data that can be integrated with toxicity-related endpoints from other databases. These include the cell-based Connectivity Map[46] and the Library of Integrated Network Cellular Signatures (LINCS).[47] Section 1.3.1 discusses all of the databases in detail.

The field of toxicogenomics can be broadly split into method-based categories: differential gene expression analysis and its pathway enrichment, compound signature matching, protein-protein interaction networks, and the use of co-expression network methods. Numerous research papers involve the integration of multiple methods.

### 1.3.1 Toxicogenomic databases

Toxicogenomic databases are strictly those containing 'omics' data with toxicity related endpoints (see DrugMatrix and Open TG-GATEs section below). However, these are complemented by orthogonal data sources that can add mechanistic or predictive aspects to a study. These include the 'Comparative Toxicogenomic Database'[48] (which includes pairwise associations between compounds, genes, pathways, and phenotypes), pathway databases, Gene Ontology lists, and transcription-gene associations. There are also large, purely transcriptional databases (Table 1-5).

#### *1.3.1.1 DrugMatrix and Open TG-GATEs*

DrugMatrix was created as a commercial database in 2006, before being brought into the public domain in 2011. It focuses on compound induced rat data, with transcriptomics data from liver, kidney, heart, and thigh muscle. Crucially, it provides histopathology, haematology and other clinically relevant data on these rats during and post exposure. It was described in detail in previous work.[44] Additionally, it contains a complete matrix for protein activities.

Open TG-GATEs[45] was published in 2012 and follows a similar protocol to DrugMatrix. The acronym stands for 'toxicogenomics project-genomics assisted toxicity evaluation system' and was created in Japan. It also contains compound-induced gene expression data and histopathology observations from rat liver and kidney. Additionally, it contains cell line data on rat and human primary hepatocytes. It should be noted that, while there are similarities with DrugMatrix, there are a few significant differences: the difference

in the dose that the rats were exposed to. DrugMatrix defines the maximum tolerated dose as ‘a 5-10% reduction in weight gain over five days of daily dosing’, and an alternative dose that was determined from literature and expert opinion. Open TG-GATEs, on the other hand, has three dose levels. The highest is defined as that which induces the ‘minimum toxic effect over the course of a 4 weeks toxicity study’. The other two are 30% and 10% of this level. This level sets a lower dose, which reflects the difference in chemical spaces that these databases cover. DrugMatrix selected its compounds to reflect a wide area of therapeutic chemical space. Open TG-GATEs selected compounds that had previously been annotated as nephro- or hepato- toxic. Both are summarized in Table 1-3.

These databases form the basis for hypothesis generation and validation in this work, and their data is reviewed in detail in Chapter 2.

*Table 1-3 Summary of DrugMatrix and Open TG-GATEs databases. Both contain in vivo gene expression data and histopathology observations, and so are suited for this study. FED is fully effective dose (expert defined), MTD is maximum tolerated dose (5-10% weight loss over a 5-day daily dosing study), ‘highest dose’ is defined as ‘minimum toxic effect over the course of a 4 weeks toxicity study.*

| | DrugMatrix | Open TG-GATEs |
|---|---|---|
| Source | Iconix/National Toxicology Program (NTP) | Japanese National Toxicogenomics Project |
| Number of compounds | ~600 | 170 (156 public compounds) |
| Gene expression Organ data | 4 (heart, liver, thigh muscle and kidney) | 2 (liver and kidney), plus cell lines (primary human and rat hepatocytes) |
| Time points (days) | 0.25, 1, 5, 7,14 | 0.25, 1, 4, 8, 15, 29 |
| Dose type | Single and repeat dose | Single and repeat dose |
| Dose level | 2 levels (FED and MTD) | 3 levels (ratio of 1:3:10 where 10 is ‘highest dose’) |

*1.3.1.2 The Comparative Toxicogenomics Database (CTD)*

The Comparative Toxicogenomics Database (CTD) is a publicly available database, first created in 2004.[49] It aims to understand the effect that environmental exposures have on human health and consists of manually curated associations between chemicals, genes, diseases, and phenotypes as well as inferred interactions. There are monthly data releases that update its known associations (ctdbase.org). The current (as of August 2019) database's contents are shown in Table 1-4.

*Table 1-4 Curated associations within the Comparative Toxicogenomics Database (August 2019)*

| Curated Interaction type | Counts |
|---|---|
| Chemical – Gene | 1,972,064 |
| Gene - Disease | 39,276 |
| Chemical - Disease | 216,163 |
| Phenotype based interactions | 208,058 |

In addition to these manually curated associations, the CTD also infers associations between its constituents. This is based on the hypothesis that "Chemical A is associated by inference with Disease B because Chemical A has a curated interaction with Gene C, and Gene C has a curated association with Disease B". This increases the number of gene - disease associations by 26,240,322 and the number of chemical – disease associations by 2,457,395.

*1.3.1.3 Transcriptomics databases*

Transcriptomics databases may also be used in the determination of toxic mechanisms of action, and so are applicable to this work. There are two groups of databases: those under the same/very similar experimental conditions (Connectivity Map (CMap) and LINCS), and those that collect a variety of gene expression data from individual scientific studies (Gene Expression Omnibus (GEO)[50] and ArrayExpress[51]). These are summarised in Table 1-5 CMap and LINCS are related databases that originated with CMap in 2006, with the gene expression profiles of 164 small-molecule compounds. This was upgraded to 1,309 drugs across five cell lines.[46] To follow up on this work and to widen the chemical space of the

xenobiotics, LINCS was created. Whilst continually updated, it contains the gene expression profiles of over 20,000 compounds on up to 77 cell lines, measured on the L1000 chip. This chip measures 978 so called landmark genes (selected to represent the greatest proportion in gene expression variation) and predicts the remainder of the genome-wide expression levels. However, there are concerns on the internal reliability and replicability of this approach.[52,53]

GEO is a repository of gene expression data, taken and uploaded from a wide range of experiments and so varies considerably with species, *in vivo/in vitro*, dose, and whether the profile is that of disease or compound-induced.[50] ArrayExpress is almost a sub-database of GEO, where the gene expression data have been curated (there are a few unique experiments).[54]

*Table 1-5 Summary of public transcriptomic databases*

| Database | Source of gene expression data | Samples | Link |
|---|---|---|---|
| Connectivity Map | 5 human cell lines | 6,400 | http://clue.io/cmap |
| LINCS | 77 human cell lines | 1,328,098[55] | http://www.lincsproject.org/ |
| GEO | 16 species, variety of cell lines | 3,167,476 | https://www.ncbi.nlm.nih.gov/geo/ |
| ArrayExpress | 16 species, variety of cell lines | 2,381,203 | https://www.ebi.ac.uk/arrayexpress/ |

*1.3.1.4 Pathway databases*

A biological pathway is a set of reactions that lead to the creation of a certain product or change in the cell. It can include altering the expression of genes or the creation of new proteins or fats.[56] Within gene expression, pathways are considered as the set of genes that act together to perform a biological function. As such, there are multiple databases that bring together these gene sets, termed pathways. The definitions between databases change and may even be considered as somewhat arbitrary.[57] As pathways consist of genes, they are

species specific, and so care must be taken in their use. Commonly used pathway databases are summarised in Table 1-6. When considering rat models, the 'Rat Genome Database' provides the relevant pathways.[58]

*Table 1-6 Commonly used pathway databases. Table adapted from Alexander-Dann et al.*[59]

| Database | Description | Comment | Link |
|---|---|---|---|
| WikiPathways[60] | Integrated collection of different pathway databases | Freely available, everyone can curate | https://www.wikipathways.org/ |
| Reactome[61] | Large database with a focus on signaling pathways | Free and the largest database of its kind | https://reactome.org/ |
| Gene Ontology[62] Reviewed:[63] | Gene product functional annotation in a hierarchically structured ontology | Contains annotations at multiple levels of specificity | http://www.geneontology.org/ |
| Kyoto Encyclopedia of Genes and Genomes[64] | One of the oldest pathway databases; content constantly updated | Very good metabolic pathway collection, but became partly paid for use and at some parts the curation is arbitrary | http://www.genome.jp/kegg/ |
| Ingenuity Pathway Analysis[65] | A complete user-friendly pathway analysis tool, which is even capable to predict causal relationships | Capable of sophisticated analysis, commercial | https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/ |
| Molecular Signature Database[66] | The Broad Institute's pathway signature collection | Different molecular signatures can be determined according to user, easy compatibility with GSEA | http://software.broadinstitute.org/gsea/msigdb |

### 1.3.2 Methods used in the field of toxicogenomics

The previous section focused on the different databases that find application in the field. The following highlights their usage and current developments. These are roughly split via

method type, although successful research does integrate many types. Herein, determining differentially expressed genes (DEGs), pathway enrichment, and co-expression network methods shall be discussed. These methods are not solely used in the field of toxicogenomics and have been used in the study of disease, compound mode-of-action and cellular biology.

#### *1.3.2.1 Differentially Expressed Genes (DEGs)*

Once gene expression data has been measured by microarray or RNA-seq and processed, it can be applied towards toxicity-related endpoints and anchoring phenotypes. The most common way to do this is to determine differentially expressed genes. A gene is considered differentially expressed if its observed expression is significantly different between two experimental conditions.[67] This is statistically determined using fold change (expression in one condition divided by its expression in the other condition) which is expressed in logarithmic form (usually base 2). The logarithm is to ensure linear mapping when considering the experimental conditions in reverse: for example, if a gene is expressed four times more in one case, its fold change equals 4. The reverse of that (i.e. underexpressed) would be 0.25. Extending this means anything overexpressed is mapped between one (no change) and infinity (hypothetically), whereas anything underexpressed is mapped to between zero and one. This non-linear mapping can lead to biased weights in gene expression values' importance. A corrected t-test can then be used to determine the false discovery rate (and so statistical significance).[68]

Separately to the fold change method, methods that utilise a gene's entire expression distribution have been developed (e.g. Bayesian and counting methods). 'Limma' is a commonly used computational package (in R) that determines DEGs using linear modelling and Bayesian estimates of a gene's variation.[69]

There are other parametric and non-parametric tests to determine DEGs, including rank product, t-statistics and B-statistics. A study compared these methods, quantified by their ability to predict rat nephrotoxicity.[70] All the methods generated "reasonably performing classifier[s]", however t-test and fold change were the most balanced performance in specificity, accuracy, and sensitivity with 83.6%, 81.0%, and 72.2% respectively. Interpretation and enrichment of this data give insight into the roles that PPAR, RXR, and D vitamin receptor play in the tubule toxicity pathways. This is illustrative of the importance that compound-induced gene expression data can play when determining toxic mechanisms of action.

There are limitations with solely determining differentially expressed genes: investigations suffer from high dimensionality (a large number of genes are being measured, but with a small number of repeats) and noise which masks the compound-induced signal. The latter is important if a compound does not have a large transcriptional level effect, so the gene expression profile is dominated by noise.[71,72] Hence, integrating this DEGs with known biological pathways may help to deconvolute the signal.

*1.3.2.2 Pathway enrichment*

Pathway enrichment is split into two categories: functional class scoring (FCS) and over-representation analysis (ORA).[73] Both methods require gene sets for their enrichment, as reviewed in Section 3.1.1.4 and Table 1-6.

FSC scoring is based on assessing the expression changes of genes in a list (gene set) between different experiments (i.e. treated vs. control). Its most common method is gene set enrichment analysis (GSEA).[74] GSEA was used in investigations into polycyclic aromatic hydrocarbons in rat liver.[75] It suggested the involvement of PPAR signaling pathway within the liver.

Over-representation analysis determines whether a pathway (gene set) is overrepresented in a list of genes (usually DEGs). This is typically performed with a Fisher exact test, hypergeometric distribution test or Jaccard Index. A one-sided Fisher exact test is the same as the hypergeometric distribution test, with the null hypothesis being that there is no significant association of the gene set with the DEGs. The chances of this occurring are calculated using Table 1-7.

*Table 1-7 2x2 confusion matrix for Fisher Exact Test*

| | DEG | All other genes |
|---|---|---|
| Genes in pathway | a | b |
| All genes not in pathway | c | d |

The calculation for obtaining such a set is:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{a+b+c+d}{b+d}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,(a+b+c+d)!}$$

Where $\binom{n}{k}$ is the binomial coefficient and ! is factorial. The p-value here represents the probability that this association would occur by chance. Often, harsh p-value cutoffs are used to determine statistical significance. Pathway databases may contain thousands of gene sets and so using the above calculation alone would result in an unacceptable number of false positives (e.g. if using a cutoff of p = 0.05 to determine significance and 1,000 pathways being tested, 50 pathways would be incorrectly labelled as significantly associated to the DEGs. To get around this problem, many 'multiple hypothesis testing' corrections are available and are summarized in Table 1-8.

*Table 1-8 Summary of multiple hypothesis testing corrections*

| Correction | Method | Comment |
|---|---|---|
| Bonferroni | Divide the computed score by the number of tests | May be very stringent |
| Benjamini-Hochberg | $P < \frac{i}{m} Q$ where p is the p value, i is the rank (of the particular p value compared to all those tested), m is the total number of tests and Q is the accepted false discovery rate | Less sensitive to overall number of tests |
| Benjamini & Yekutieli | $P < \frac{i}{m \, x \, c(m)} Q$ where c(m) allows for dependency of tests. | Incorporates the dependence of tests |

The main limitation to all pathway-based enrichments is the curation bias of the database; there is a bias towards the study of genes involved where current research is more invested e.g. genes implicated in cancer which then have more entries within pathway databases. Additionally, it cannot give entirely novel mechanisms of toxicity as the genes must be annotated with pathways for them to occur. However, such methods do put gene-level outputs into context.

*1.3.2.3 Co-expression network methods*

Gene expression experiments result in high dimensional data, as previously noted. As such, methods can make use of the entire measured transcriptome to make use of gene – gene interaction behaviour, and so provide sophisticated dimensionality reduction. Co-expression network methods use such information, based on the hypothesis that highly correlated and co-expressed genes are involved in the same biological functions. They are split into two predominant categories: data-driven or knowledge based. They have recently been reviewed here.[76]

Early work with co-expression networks was a method called "context likelihood of relatedness".[77] Here, the mutual information of genes creates a similarity network by the estimation of a gene-gene interaction pair against the full mutual information distribution per gene. This was applied to 2,4,6-trinitrotoluene (TNT) exposed human and rat hepatocytes.[78] This showed concordance between human and rat cells, an important consideration when determining suitable animal models for human toxicity.

Following on from "context likelihood of relatedness", the 'iterative signature algorithm' (ISA) was created.[79] This method results in 'modules', which are groups of highly correlating genes. The method relies on starter (or input) genes, which are thought to belong to separate modules. These are typically generated from hierarchical clustering gene expression matrices or may be randomly generated. Once created, the modules are refined by adding/removing genes iteratively to determine correlation. Gene and condition thresholds determine the modules size and stringency. Such restrictions allow for genes to be in multiple modules. However, the starter genes may mean that the final modules are local, not global, minima and so the method may not find all true modules for a given simulation.

The most prominent method is 'weighted gene co-expression network analysis' (WGNCA).[80] It has been shown to be one of the most suited methods for determining the main expression and functional effects of samples within a dataset.[76] It is non-exhaustive (i.e. genes do not have to fit into a module but can be left out) and, whilst two main parameters are required to be optimised, it is shown to be relatively (to other module determining methods) insensitive to parameter tuning. The method to determine modules consists of three main steps: (i) creation of a gene-gene correlation matrix based on gene expression signatures, (ii) weighting of co-expression, (iii) module determination. These are shown in Figure 1-4. Initially, all genes (or probes, depending in the input levels) are correlated against each other, resulting in a numerical matrix with values between zero and one (for an unsigned

network, -1 and 1 for a signed network). This is then raised to a softpower, β. This parameter is determined from the data to optimise the 'scale free' property of the network.

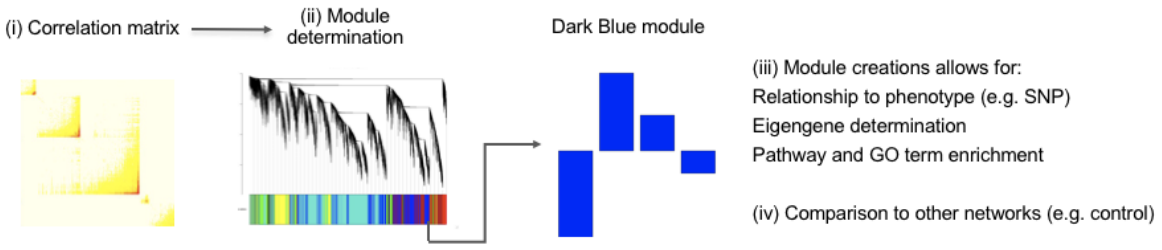



*Figure 1-4 The overview of weighted gene co-expression network analysis (WGCNA). Adapted from Alexander-Dann et al.*[59]

Scale free properties in biological networks are a controversial issue. 'Scale free' refers to the power law distribution of degree (aka some nodes in the network have many more connections than others, and the distribution of these connections follows a power law). This is modelled in Figure 1-5 where P(k) is the fraction of nodes that have k connections. It is a central claim in network research that "real world" networks, particularly complex, biological networks, obey this distribution. Whether or not the biological network truly follows this distribution, by raising the correlation network to the power of β, the signal to noise ratio is changed. The softpower is usually in the range 5 – 20, and so high correlation values (close to one) will be relatively unchanged. Low correlation values (near zero) are assumed to be noise and are reduced heavily, e.g. $0.9^{10} = 0.35$ compared to $0.3^{10} = 5.9\ x\ 10^{-6}$.

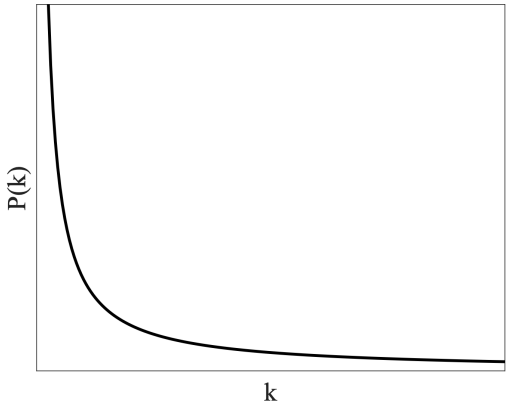



Figure 1-5 Scale free distribution of connections in complex networks. P(k) is the proportion of nodes that have k connections.

The result of this forms a weighted adjacency matrix for all genes (nodes) in the network. The next step is to use the network properties (specifically node's connectivity) to further weight the matrix into a topological overlap matrix (TOM). This is done by

determining the pairwise topological overlap, *w*, of two nodes, *i* and *j*. For biological networks, the suitable equation for this is:[81,82]

$$w_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$$

Where $l_{ij}$ is the connectivity between two nodes, $a_{ij}$ is the adjacency weighting from the above matrix, and $k_i$ is the total connectivity of one node. This results in a similarity metric which can be convert to a dissimilarity network by simply subtracting from 1. This network is complete (i.e. all genes are connected to all other genes with a non-zero weight). Hierarchical clustering of this dissimilarity matrix genes defines modules, making a non-complete network. Thus, creating the output of WGCNA.

It is the output of WGCNA that is most useful: each module represents a sub graph of the original network formed of highly connected genes which can be enriched with pathway level annotations (in the same manner as the DEG to pathway enrichment). The modules can be related to external information (e.g. phenotypes) and they may be tested to determine their conservation in other networks.[83] This has been performed in many applications, including non-toxicity related endpoints, to determine gene markers for diseases, weight phenotypes, and growth and development as endpoints.[84–87]

Co-expression methods have significant potential in understanding and predicting compound-induced toxicity. However, they are reliant on correlation arguments. Correlation is not the same as causation and so care must be taken in the reliability of the created associations. As they are indeed based on correlation, there are also requirements for a larger number of repeats (the standard of 3 repeats is not enough for robust correlation). Co-expression network methods are sensible data dimensionality reduction tools and represent the state of art.

### 1.3.3 Developments in the field of toxicogenomics

#### *1.3.3.1 Differentially expressed genes and pathway enrichment*

Determining differentially expressed genes has been used to suggest hypotheses for toxic mechanisms of action. For example, they were used to determine the response of neuroblastoma cells to exposure to MPP+ (1-methyl-4-phenyl-pyridinium, a known toxicant and a model for Parkinson's disease).[88] Here, differentially expressed genes (DEGs) were defined by their fold change (greater than 1) and the microarray was confirmed by RT-qPCR.

This list of DEGs included two transcription factors, namely c-Myc proto-oncogene and RNA-binding protein 3, suggesting their involvement in toxicity development. The same toxicant was further examined in a time-dependant manner, which lead to the determination of 79 DEGs, which passed strong significance cut-offs.[89,90] Histones, such as H2AFJ, H3F3B, HIST1H2AC, HIST1H2BD, HIST1H2BG, and HIST1H2BK, were differentially expressed and this suggested that the destabilization of nucleosomes occurs after initial exposure to the xenobiotic. These techniques help to understand the mechanisms of toxicity. Furthermore, DEGs can be used as biomarkers and/or variables in predictive models.[91–93]

It was used in the mechanistic study of crystalline silica on human lung adenocarcinoma cells (A549) and rat lungs.[94] Crucially, this study showed concordance between the two species, and it suggested novel mechanisms of silica-induced pulmonary toxicity; centred on different dual specific phosphatase (DUSP1 and DUSP5) and growth arrest proteins (GADD34 and GADD45α). The method was similarly used in melphalan-induced vascular toxicity.[95] Here, DEGs were mapped to transcription factors, with functional modules indicating MYC and NF-κB1 showing significant involvement. From these, the molecular mechanism of melphalan-induced vascular toxicity was hypothesised and five small molecules were suggested to modulate the crucial transcription factors, and so overcome the toxicity.

Further work, from DEGs and enriched pathways was performed on data from Open TG-GATEs.[96] Reactome pathways were used to construct a "*computationally predicted adverse outcome pathway*" i.e. a workflow describing all the key molecular and cellular steps, from a compound binding a molecular target all the way to the observable phenotype, for a compound with specific histopathology phenotype. It was exemplified with the mechanism of fatty liver disease, caused by exposure to carbon tetrachloride. Similarly, methods with pathway analysis and protein-protein interaction networks have been studied to determine toxic pathways in liver injuries in rat.[97,98] These studies showed decreased metabolism in the liver, with increased inflammatory pathway activity, as well as activation of fibrosis relevant genes.

Determining DEGs and pathway level enrichments forms the basis of most toxicogenomics analyses. However, the output of these may be hundreds of DEGs and pathways, with no clear way of prioritising them, so determining primary mechanisms of action is not always clear. To that end, further methods that use gene expression data have been developed, most notably, the use of co-expression network methods.

*1.3.3.2 Co-expression network methods*

ISA has been applied to both DrugMatrix and Open TG-GATEs, to determine gene expression signatures associated with 'chemical and drug-induced liver injuries'.[99] Using clinical pathology, organ weight and histopathology observations, the authors generated 25 diverse toxicity-related endpoints. Theses labels were shown to not be independent and were highly correlating. These labels segregated gene expression profiles, from which modules were created. Using ISA, hierarchical clustering, support vector machines, DEGs, and protein-protein interaction networks, different modules were created. ISA created modules with high enrichment of liver injury (taken from gene-disease relationships in the comparative toxicogenomic database). Specifically Sod2, Gulo and Car3 were associated with periportal lipid accumulation, and Obp3 and Rgn were associated with periportal fibrosis. Open TG-GATEs was used to validate this approach. Associations to acute kidney injury (AKI) have also been determined with this method.[100] The created modules were used to determine a list of 30 genes, to be used a biomarker of AKI. They were validated against randomly selected genes and those from additional AKI gene expression data from GEO. Genes known to be involved in AKI were found, including Havcr1, Clu, and Tff3. Importantly, novel gene – phenotype associations were found, namely Cb44, Plk2, Mdm2, Hnmt, Macrod1, and Gtpbp2. These were confirmed in non-compound induced AKI models, implying a non-specific response to injury.

Toxicity relevant studies have also been performed with WGCNA. The dose-dependent carcinogenic effect of chloroprene in mice was investigated.[101] Two modules were shown to be important in differentiating the outcome, with seven hub genes (genes with the highest number of connections, quantified by degree) found critical for carcinogenesis of lung tissue.

Numerous studies have used WGCNA in application for rat liver toxicity. Using a toxicity label based on the histopathology observation of 'liver periportal fibrosis' (with grade of minimal or above), relevant modules were determined.[97] This was based on DrugMatrix data and found known, relevant genes such as TIMP1, APOA1, CTGF, LGALS3, TGFB1, and MMP-2, within a module annotated with 'liver cirrhosis' from the CTD (N.B. liver fibrosis is not a curated term) and the 'extracellular matrix organization' GO term. Novel associations were also determined: LGMN, a cysteine protease that functions in ECM remodeling, and PLIN3, which is known to be involved in the pathogenesis of steatosis. Interestingly and importantly in predictive toxicology, gene expression profiles of known

toxicants were linked to the phenotype before it was visible, showing the potential for early-stage biomarkers.

Sutherland *et al.* also used WGCNA to associate gene expression data with histopathology from DrugMatrix and Open TG-GATEs.[102] They created one large network across all different compound-induced gene expression data and determined 415 specific modules. They then enriched these modules with particular histopathology observations based off their relevant correlation with a module's eigengene (the dimensional reduction of a module to a vector with a value for each sample). This generated a large number of phenotype-gene associations, both novel and established. They suggested a mechanism of hepatotoxicity involving endoplasmic reticulum stress and Nrf2 activation.

### 1.3.4 Limitations of the field of toxicogenomics

Despite the advancement and progress in both methods and available data, there are numerous issues facing the field of toxicogenomics. Most notably, the evidence generated in the studies has not yet achieved regulatory approval for a new chemical entity.[103] However, they are being used for the 'weight-of-evidence' evaluations for substances' mode-of-action. The reliability, reproducibility and efficacy are crucial for incorporation into both drug development and regulatory approval. This paves the way for the work in this thesis to be presented.

## 1.4 Overview of work contained in thesis

This work presents an analysis of DrugMatrix and Open TG-GATEs across different data domains: chemical structure, physicochemical descriptors, histopathology and gene expression spaces. From these, the creation of histopathology signatures is performed to generate toxic groups (Chapter 2).

These groups are then analysed through the use of gene expression data (via co-expression networks) to determine gene-phenotype associations at the time point of the phenotype (Chapter 3) and preceding time points (Chapter 4).

# 2. Evaluation of data domains to classify toxic classes in DrugMatrix and Open TG-GATEs

DrugMatrix and Open TG-GATEs are two large, public databases that contain data from various different domains: chemical structure, gene expression, histopathology, bioactivity, and clinical chemistry. With the wealth of data available, it is vital to determine which data is toxically relevant, within the domains of all the measured limitations. Herein, these domains are assessed with a view to determine data driven toxic groups.

First, chemical structure and physiochemical properties are analysed. The compounds in the toxicogenomics databases represent at range of chemical structures, and so are compared to ChEMBL[104] and DrugBank.[105,106] used to represent known chemical space and known therapeutic spaces, respectively. ChEMBL contains 1,879,206 million compounds, with over 15 million activities. DrugBank contains 13,345 chemical entitites, of which ~2,000 are approved small molecule drugs and ~6,000 are experimental small molecules. As such, these make ideal comparisons to DrugMatrix and Open TG-GATEs, on a chemical structure and physico-chemical property level.

Second, the effect of primarily time point and secondarily of dose on gene expression is considered. Data exploration of the gene expression data available is performed, tied to time point, and related to histopathology observations.

Third, histopathology data from DrugMatrix and Open TG-GATEs are analysed to determine their (in)dependence for the classification of toxic groups. Due to differences in nomenclature used, terms are mapped to the histopathology ontology HPATH. This was developed by the eTox consortium for the standardisation of histopathology terms between experiments.

The resulting toxic groups were investigated with respect to their similarities and differences in terms of compounds, doses, time points and histopathology observations.

## 2.1 Methods

### 2.1.1 Chemical space

The compounds selected for DrugMatrix and Open TG-GATEs were compared to each other and to ChEMBL (version 24)[104] and DrugBank (version 5.0.1)[105], using structural fingerprints and physico-chemical descriptors. Standardised SMILES were downloaded from each database, and processed in KNIME (version 3.7.2).[107] Physico-chemical descriptors

were calculated for all of DrugMatrix and Open TG-GATEs, and randomly selected subsets of 50,000 compounds from ChEMBL and DrugBank, using the RDKit 2D "calculated descriptors" node. The descriptors are shown in Table 2-1.[108] Principle component analysis (PCA) was then performed to compare the variation of chemical properties between databases, using the 'fviz' function on normalised data from the 'factoextra' package[109] in R, version 3.5.1.[110]

The chemical structures in each of the databases were compared using 'extended connectivity fingerprints' (ECFP), a circular, topological fingerprint, with radius = 4. The Tanimoto similarity of ECFP4 fingerprints between every compound in each database and its 5 nearest neighbours in all databases was calculated using Python (version 3), in the same method as previous papers.[111–113] In short, SMILES were converted to binary fingerprints. Tanimoto similarity is the determined using the overlap of matching bits. The Mann-Whitney, Kolmogorov–Smirnov and t-tests were then performed to determine the significance of the resultant distribution using the 'stats' package in R.[110]

*Table 2-1 2D physico-chemical descriptors calculated used RDKit that build the basis for further analysis*

| 2D descriptor | Description |
|---|---|
| SlogP | logP (partition coefficient) using surface area contributions |
| SMR | molecular refractivity |
| LabuteASA | Labute's Approximate Surface Area |
| TPSA | topological polar surface area |
| AMW | atomic molecular weight |
| ExactMW | exact molecular weight |
| NumLipinskiHBA | number of Lipinski hydrogen bond acceptors |
| NumLipinskiHBD | number of Lipinski hydrogen bond donors |
| NumRotatableBonds | number of rotatable bonds |
| NumHBD | number of hydrogen bond donors |
| NumHBA | number of hydrogen bond acceptors |
| NumAmideBonds | number of amide bonds |
| NumHeteroAtoms | number of heteroatoms |
| NumHeavyAtoms | number of heavy atoms |
| NumAtoms | number of atoms |
| NumRings | number of rings |
| NumAromaticRings | number of aromatic rings |
| NumSaturatedRings | number of saturated rings |
| NumAliphaticRings | number of aliphatic rings |
| NumAromaticHeterocycles | number of aromatic heterocycles |

| | |
|---|---|
| NumSaturatedHeterocycles | number of saturated heterocycles |
| NumAliphaticHeterocycles | number of aliphatic heterocycles |
| NumAromaticCarbocycles | number of aromatic carbocycles |
| NumSaturatedCarbocycles | number of saturated carbocycles |
| NumAliphaticCarbocycles | number of aliphatic carbocycles |
| FractionCSP3 | fraction of sp3 hybridised carbon atoms |
| Chi[1v-4v] and Chi[1n-4n] | graphical parameters representing the "complexity" of a molecule |
| HallKierAlpha | Hall and Kier Alpha value (representing connectivity of a molecule) |
| kappa[1-3] | Hall and kier Kappa values |
| slogp_VSA[1..12] | logP using virtual surface area (subdivided surface area) |
| smr_VSA[1..10] | molecular refractivity (subdivided surface area) |
| peoe_VSA[1..14] | total partial charges of heavy atoms |
| MQN[1..42] | molecular quantum numbers |

### 2.1.2 Gene expression methods

The DrugMatrix[114] data was accessed *via* Gene Expression Omnibus (GSE57822).[115] Only Affymetrix whole genome GeneChip Rat Genome 230 2.0 Array data, from liver at 1 and 5 days of daily dosing was selected using R, as these time points had the highest amount of histopathology observations (477 individual histopathology observations at 5 day and 188 at 1 day (the next highest)).[110] This process gave a total of 1004 CEL files, covering 129 unique compounds and 159 compound-dose instances which also possessed histopathology data in DrugMatrix. Compound-dose pairs were considered as separate instances as they result in different histopathology profiles. Open TG-GATEs[45] was accessed *via* ArrayExpress[51] (E-MTAB-800) and liver data from the 1 and 4 day stage of daily dosing were selected, giving a total of 714 CEL files, covering 67 compounds and 109 compound-dose instances.

All CEL files for DrugMatrix and Open TG-GATEs were pre-processed using the Robust Microarray analysis (RMA)[116] functions and quartile normalisation in the 'affy' package in R (version 1.62).[117] Differentially expressed genes were determined using the 'limma' package[69] as follows: Linear models were fitted (lmfit) to the distribution of probe intensities for each compound -dose-time point instance and controls. A moderated T-test was performed to determine the significance of the overlap of the distribution. The cut-off for the Benjamini-Hochberg corrected p-value was 0.05 and the $\log_2$(fold change) cut-off was 2. Probes were matched to genes using 'AnnotationDbi' and 'Rat2302.db'.[118] When multiple

probes matched to a single gene , the probe with the smallest p-value was used.[119] When a probe matched to multiple genes, the log fold change and the p-value from the probe was used for each.

Principle component analysis (PCA) was then performed to compare the variation of chemical properties between databases, using the 'fviz' function on the RMA adjusted data from the 'factoextra' package[109] in R, version 3.5.1.[110] Multidimensional scaling was performed using the 'cmdscale' function.

To determine the consistency of compound induced gene expression, each compound-dose-time-point instance was correlated internally (against each biological repeat) and externally (against other compound-dose-time-points). The RMA pre-processed gene expression data from DrugMatrix was linearly correlated (Pearson's product-moment coefficient) in R, using the base package. [110] Gaussian kernel smoothing was used for ease of visualization, within the same R package.[110] The two distributions can be seen in Figure 2-1. A Kolmogorov–Smirnov test (KS) test was applied to determine if the distributions were significantly different.

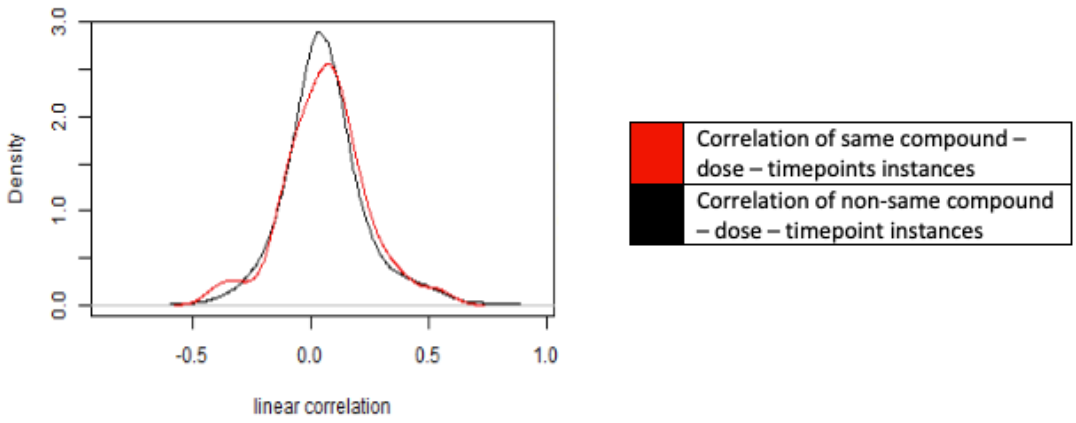



*Figure 2-1 Gaussian kernel density plots showing the correlation of gene expression data, when correlations are calculated for the same compound-dose-timepoint (red, n = 109) and not (black, n = 187889) for the DrugMatrix database. A KS test was performed, to test for the independence of each data. The null hypothesis (that both samples are from a larger distribution) was rejected with a p value = 0.0456, and the test statistic, D = 0.130. This means that whilst the size effect is small, compound-dose-timepoint gene expression profiles are more internally correlated.*

The results of this showed a small (size) but statistically significant differentiation between instances with the same compound and dose, to those without. This shows that there is a compound, dose, and time point signal within the gene expression data.

### 2.1.3 Histopathology methods

Histopathology data were downloaded from DrugMatrix.[120] Each histopathology observation was mapped to the Histopathology Ontology, HPATH, which was created by the eTox consortium.[27] The full list of mappings can be seen in Supplementary Table 1. Histopathology data for Open TG-GATEs were downloaded.[121] In this case no mapping to HPATH could be performed due to the lack of meta data for each observation. However, matches of histopathology terms between DrugMatrix and Open TG-GATEs could still be determined by manual comparison.

The histopathology data distributions over organ, time point, database and severity were plotted in Knime and R. Histopathology observations were treated as vectors, and their independence were tested using the Fisher's exact test. The p value was corrected using the Benjamini-Hochberg correction. This was performed using the 'stats' package in R.

In order to consider all measured histopathology readouts simultaneously histopathology signatures were created, which is represented in Figure 2-2. The histopathological assays from DrugMatrix mapped to HPATH were binarized using a cut-off ≥ 1 (slight) (see Figure 2-2 for the severity score levels) for each rat, as in previous studies.[97] These were summarized per compound-dose condition as a signature and identical signatures (in binary space) grouped together. This created sets of compound-dose instances at the 5 day stage that had identical observed histopathology profiles (in binary representation of readouts). While the time element of readouts is also scientifically relevant, this is considered in Chapter 3. 86 signatures were determined in total, however only 6 signatures consisted of more than 4 compound-dose instances. Histopathology signatures were created from Open TG-GATEs directly, without mapping to HPATH due to a lack of available information on the observations in English. 30 signatures were determined in total, 20 of which were larger than 4 compound-dose instances (Supplementary Table 2).

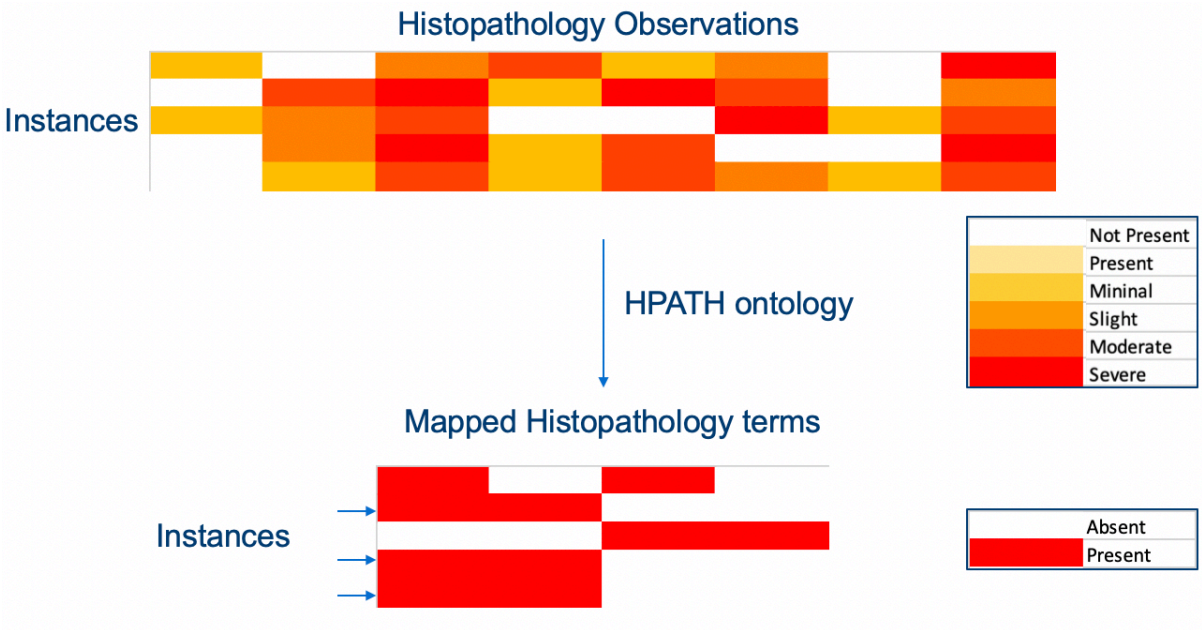



*Figure 2-2 Visualisation of the histopathology mapping to derive a histopathology signature. The heatmap above represents the data as available in DrugMatrix, using a scoring scheme from 0-5 (observation not present to severe). Each histopathology assay was mapped to a base HPATH term, resulting in many-to-one mappings or near synonyms. The score was binarized using a cut-off of minimal or above. The number of histopathology observations is therefore reduced and, as a result, more compound-dose instances could be mapped to the same histopathological signature (as indicated by the blue arrows).*

## 2.2 Results and discussion

### 2.2.1 Chemical space

DrugMatrix contains around 600 unique compounds and Open TG-GATEs contains 170 (of which 154 have publicly available structures). It has already been mentioned that the choice here reflects two different perspectives: a wide selection of therapeutic chemical space was selected for DrugMatrix, whereas Open TG-GATEs contains compounds that were previously annotated with either hepatoxicity or nephrotoxicity.

The full list of compounds in each are readily available from their respective sources.[120,121] There are 104 unique therapeutic classes in DrugMatrix, as shown in Table 2-2, out of a comprehensive list of 120 (according to the 2007 DrugMatrix Calculations White paper).[120] From this, the wide coverage of therapeutic space can observed.

*Table 2-2 Therapeutic classes of DrugMatrix compounds, showing the wide therapeutic space from which compounds were selected for DrugMatrix.*

| Therapeutic class of DrugMatrix Compounds | |
|---|---|
| Acidifying Agents | Corticosteroids |
| Acne Preparations | Corticosteroids, Systemic |
| Alkalinizing Agents | Corticosteroids, Topical |
| Amino acid metabolism Disorders | Dental Agents |
| Analgesics, Non-Opioid | Diuretics |
| Analgesics, Opioid | Drug Delivery Systems |
| Analgesics, Topical | Drug Dependence Therapy |
| Anorexiants | General Anesthetics, Inhaled |
| Antianginals | General Anesthetics, Intravenous |
| Antiarrhythmic Agents | Gonadal Hormones |
| Antibacterials | Gout-Related Agents |
| Antibacterials, Systemic | Hematopoietic Agents |
| Antibacterials, Topical | Hemorrheologic Agents |
| Anticoagulants | Hemostatic Agents |
| Antidepressants | Hypolipidemic Agents |
| Antidiabetic Agents | Hypothalamic Hormones |
| Antidiarrhea Agents | Immunomodulants |
| Antidotes | Immunostimulants |
| Antiemetics | Immunosuppressants |
| Antiepileptics / Anticonvulsants | Insecticides / Pesticides |

| | |
|---|---|
| Antifungals | Irrigating Solutions |
| Antifungals, Systemic | Laxatives |
| Antifungals, Topical | Local Anesthetics |
| Antiglaucoma Agents | Miscellaneous Anti-Asthmatic Agents |
| Antihistamines | Miscellaneous Anti-Inflammatory Agents |
| Antihypertensive Agents | Miscellaneous Cardiac Agents |
| Antimanic Agents | Miscellaneous Dermatological Agents |
| Antimigraines | Miscellaneous Gastrointestinal Agents |
| Antineoplastics | Miscellaneous Ophthamological Agents |
| Antiparasitics | Movement Disorders |
| Antiplatelets | Mydriatics |
| Antipruritic Agents | Neurodegenerative Disorders |
| Antipsychotics | Nonsteroidal Anti-Inflammatories (NSAIDs) |
| Antirheumatic Disease Modifying Agents | Opiate Antagonists |
| Antiseptics / Disinfectants | Oxytocic |
| Antispasmodic Agents | Parasympatholytic Agents |
| Antitussives, Expectorants and Mucolytic Agents | Parasympathomimetic Agents |
| Antiulcers | Pituitary Hormones |
| Antivirals | Prokinetic Agents |
| Antivirals, Systemic | Psoralens |
| Antivirals, Topical | Radioactive Agents |
| Anxiolytics | Respiratory Stimulants |
| Appetite Stimulants | Sedatives/Hypnotics |
| Astringents | Skeletal Muscle Relaxants |
| Bone Mineral Homeostasis | Sympatholytic Agents |
| Bronchodilators | Sympathomimetic Agents |
| Central Stimulants | Thrombolytic Agents |
| Cerebral Vascular Disorders | Thyroid and Antithyroid Agents |
| Chelating Agents / Heavy Metal Antagonists | Uricosuric Agents |
| Chemoprophylactics / Preventives / Protectives | Vasodilators |
| Congestive Heart Failure | Vasopressors |
| Contraceptive Hormones | Vitamins / Minerals / Nutrients |

As Open TG-GATEs does not solely contain therapeutics, such annotations are not available. However, its selection criteria of nephro- or hepato- toxic annotation limits its coverage of therapeutic space. To compare the chemical space, physico-chemical descriptors and compound descriptors from both databases and two external databases (ChEMBL and DrugBank) are compared. It is expected that DrugMatrix and Open TG-GATEs cover

different chemical subspace to each other, and less varied chemical space compared to ChEMBL and DrugBank. ChEMBL and DrugBank are larger databases and the likelihood of finding more similar compounds is higher. To combat this, a subsample of 50,000 of each is used.

Physico-chemical descriptors are used to capture the structural and physical properties of compounds. Figure 2-3 shows the result of the PCA analysis. The first three dimensions cover 57.5% of the total variance within the dataset. The subsets of ChEMBL and DrugBank cover DrugMatrix and Open TG-GATEs, indicate that the latter two do not cover novel chemical descriptor space. This indication is not proof as the total variance captured was 57.5%.

Structural comparisons of the databases show the diversity of both DrugMatrix and Open TG-GATEs. Each set of box plots in Figure 2-4 represent the tanimoto similarity the five most similar compounds in each of the databases, measured against all (or, in the case of ChEMBL and DrugBank, a subset) the compounds of the database of interest. In each case, ChEMBL and DrugBank have higher average similarity (~0.5 and ~0.4). This may reflect the larger size of these databases; each contain 50,000 compounds and so there is a higher chance of more similar compounds being measured. The dissimilarity of DrugMatrix and Open TG-GATEs (~0.3 and ~0.2 respectively) shows that both DrugMatrix and Open TG-GATEs contain a wide variety of chemical structural space.

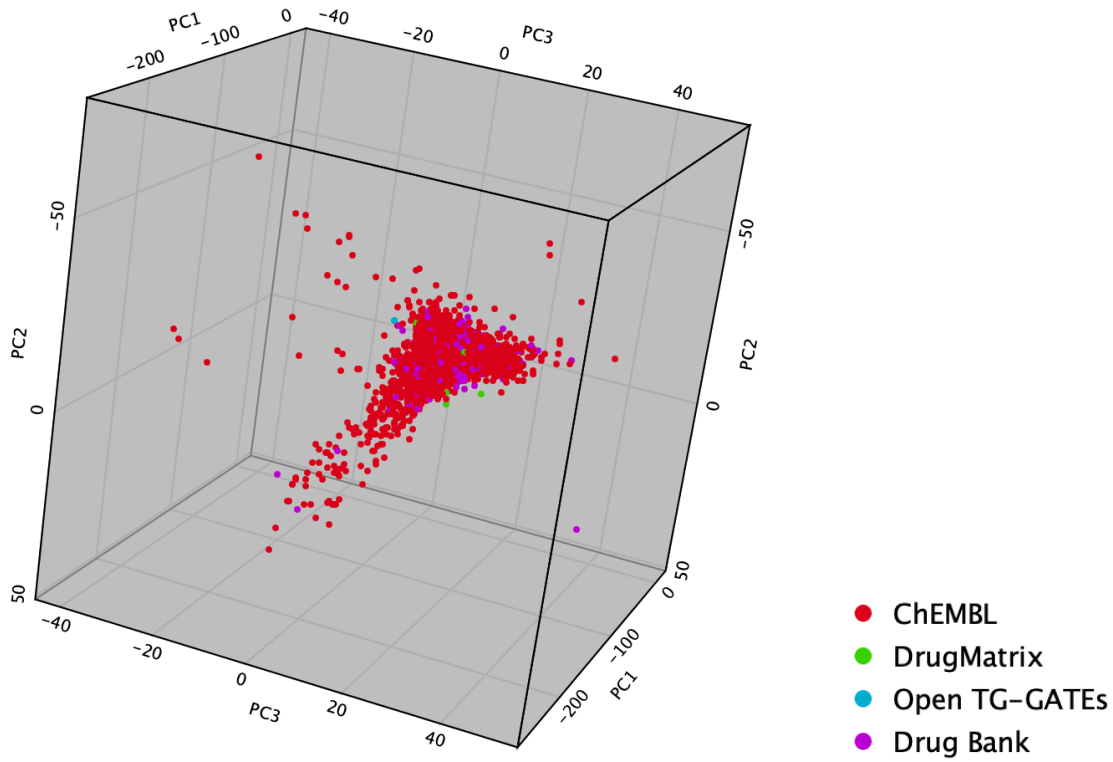



*Figure 2-3 Principle Component Analysis (PCA) of the physico-chemical descriptors of the compounds in DrugMatrix, Open TG-GATEs, and a random subsection of ChEMBL and DrugBank. Each compound from the databases projected onto the first three dimensions. DrugMatrix and Open TG-GATEs are within the areas of ChEMBL and DrugBank, reflecting a less varied chemical space.*

To determine whether DrugMatrix and Open TG-GATEs contain diverse compound structure, the distributions of their five most similar compounds were tested using three methods: Mann-Whitney, t-test and the Kolmogorov–Smirnov (KS) test. The results of these are shown in

*Table 2-3*. The distributions for similarity of the 5 closest neighbours were found to be significantly different in every case, except between DrugBank and ChEMBL. As DrugBank is a subset is a subset of ChEMBL (although the random sub-samples do not overlap here). The subsetting is to account for the difference in size of database. The T-test assumes a normal distribution of the test distributions, and these are plotted in Supplementary Figure 1.

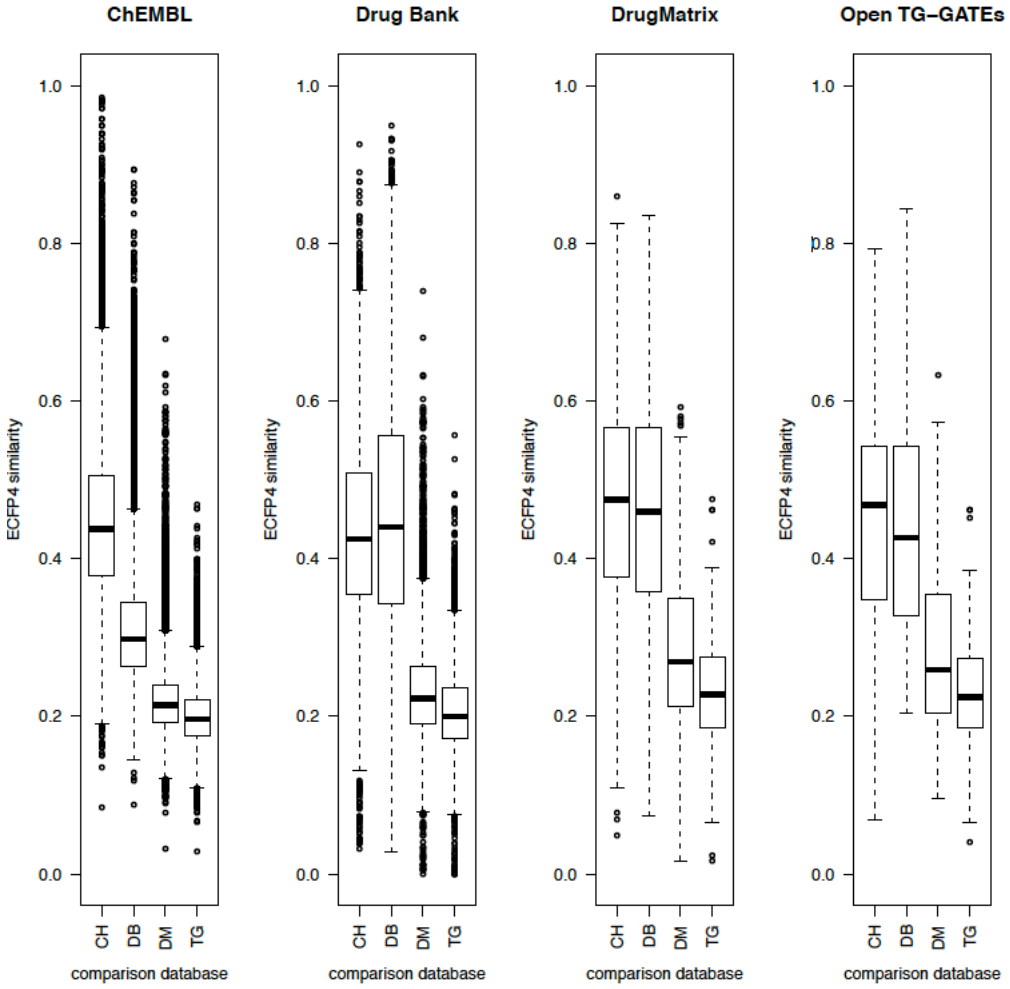



*Figure 2-4 Distributions of ECFP4 similarity (Tanimoto) for each database. Across all boxplots, ChEMBL and DrugBank show higher amounts of similarity between compounds. DrugMatrix and Open TG-GATEs show low levels of similarity within themselves, showing the diversity in their structures. CH = ChEMBL, DB = DrugBank, DM = DrugMatrix, TG = Open TG-GATEs*

*Table 2-3 Results of chemical structural similarity distribution testing. 0 represents a p value less that 1.6E-180. KS is the Kolmogorov–Smirnov test. This analysis helps to visualise the boxplots in Figure 2-4 to determine the significance between the distributions. It quantises that when using the DrugMatrix and Open TG-GATEs compounds, there is no observed difference between ChEMBL and DrugBank. DrugMatrix and Open TG-GATEs are consistently significantly distinct from each other, representing their different selection*

*criteria. Such differences mean each are outside the applicability domain of models built one the other, in chemical space.*

| | | | Test (p value) | | |
|---|---|---|---|---|---|
| **Compound database** | **distribution comparison databases** | | **Mann-Whitney** | **KS** | **t** |
| **ChEMBL** | ChEMBL | DrugBank | 0 | 0 | 0 |
| | ChEMBL | DrugMatrix | 0 | 0 | 0 |
| | ChEMBL | Open TG-GATEs | 0 | 0 | 0 |
| | DrugBank | DrugMatrix | 0 | 0 | 0 |
| | DrugBank | Open TG-GATEs | 0 | 0 | 0 |
| | DrugMatrix | Open TG-GATEs | 0 | 0 | 0 |
| **DrugBank** | ChEMBL | DrugBank | 1.70E-11 | 0 | 4.09E-24 |
| | ChEMBL | DrugMatrix | 0 | 0 | 0 |
| | ChEMBL | Open TG-GATEs | 0 | 0 | 0 |
| | DrugBank | DrugMatrix | 0 | 0 | 0 |
| | DrugBank | Open TG-GATEs | 0 | 0 | 0 |
| | DrugMatrix | Open TG-GATEs | 2.06E-138 | 0 | 1.74E-146 |
| **DrugMatrix** | ChEMBL | DrugBank | 4.15E-01 | 4.32E-01 | 3.89E-01 |
| | ChEMBL | DrugMatrix | 5.63E-53 | 0 | 2.16E-60 |
| | ChEMBL | Open TG-GATEs | 6.59E-82 | 0 | 3.42E-98 |
| | DrugBank | DrugMatrix | 6.89E-48 | 0 | 9.07E-55 |
| | DrugBank | Open TG-GATEs | 1.54E-77 | 0 | 7.64E-92 |
| | DrugMatrix | Open TG-GATEs | 8.20E-11 | 4.16E-08 | 2.73E-14 |
| **Open TG-GATEs** | ChEMBL | DrugBank | 2.19E-01 | 2.03E-01 | 2.83E-01 |
| | ChEMBL | DrugMatrix | 3.00E-28 | 0 | 2.21E-32 |
| | ChEMBL | Open TG-GATEs | 1.05E-40 | 0 | 2.86E-48 |
| | DrugBank | DrugMatrix | 2.51E-24 | 0 | 4.24E-27 |
| | DrugBank | Open TG-GATEs | 1.18E-37 | 0 | 2.62E-42 |
| | DrugMatrix | Open TG-GATEs | 2.88E-05 | 5.48E-04 | 1.67E-06 |

Compound structures and their physicochemical descriptors have been used for the prediction of toxicity (and more specifically drug-induced liver injury)[122], and when in large enough sample sizes and closely defined endpoints, this has achieved high levels of success,

with that study showing correct classification rates up to 89%. However, more complicated end points are no so easily modelled. Compound structure does not correlate well with systems level data – for gene expression, for example, compounds with a Tanimoto similarity greater that 0.85 only have a one in five chance of a similar gene expression profile when under identical conditions.[123]

### 2.2.2 Gene expression results

The effects of time, dose and database on gene expression data was determined using principle component analysis, multi-dimensional scaling and hierarchical clustering, which are shown in Figure 2-5, Figure 2-6 and Figure 2-7 respectively. In the first instance (Figure 2-5) very little can be concluded about the time dependency of DrugMatrix as only 24.4% of the variance is covered by the two dimensions. For Open TG-GATEs, 75.5% of the variance is covered and no distinguishable clusters of time and dose (low, medium and high) were formed. From this, it can be concluded that the individual compound-dose instance plays a more significant role than dose and time. Therefore, grouping of compound-dose instances to form a toxic set must be taken on data domain that is not purely gene expression data.

To determine the differences between the databases, the multidimensional scaling (Figure 2-6) shows DrugMatrix and Open TG-GATEs distinction. This result was copied in the hierarchical clustering of instances (*Figure 2-7*) (using Euclidean distance). Additionally, this showed that, within a database, instances were a similar distance from each other (they merged at heights 30-50). This is promising as, within a database, it shows that the instances have a similar behaviour (in relationship to each other). This allows hypotheses that are developed in one to be tested in another.


*Figure 2-5 Principle component analysis (PCA) of the DrugMatrix and Open TG-GATEs gene expression data at 1 and 4/5 days to show the effect of time and dose (for Open TG-GATEs). A) PCA of DrugMatrix showing the overlap of 1 and 5 day data, 24.4% of variance was covered the two dimensions shown and so little can be concluded in this plot. B) PCA of Open TG-GATEs at 3 dose levels and 2 time points, showing a lack of distinction between variables, 75.5% of variance was covered the two dimensions shown. No obvious clusters can be seen. DM = DrugMatrix, TG = Open TG-GATEs.*

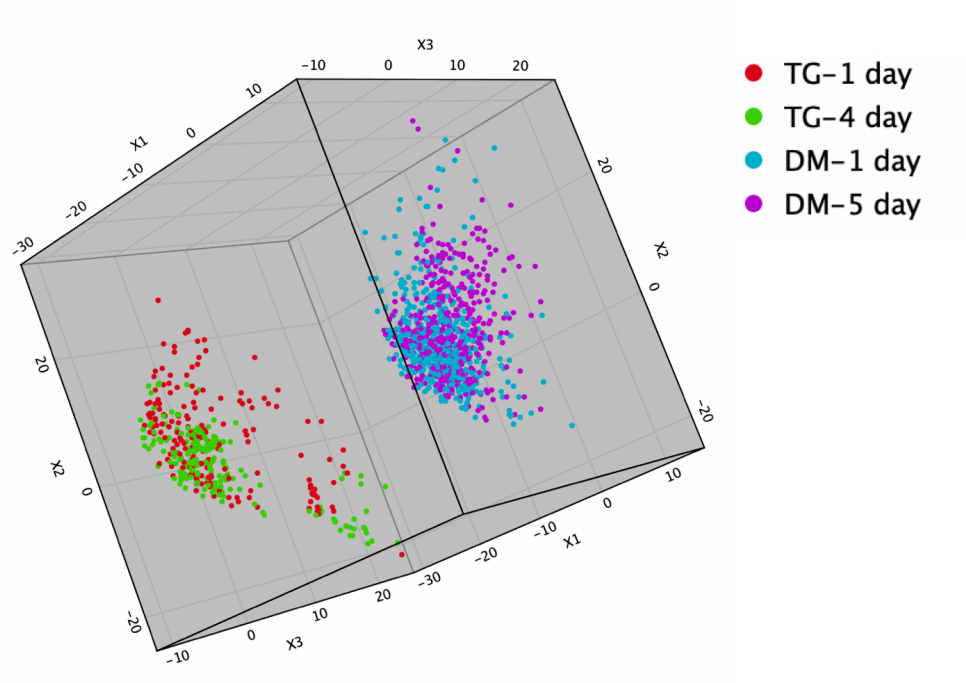



*Figure 2-6 Multi-dimensional scaling of DrugMatrix and Open TG-GATEs gene expression data, showing the database differentiation. From this, differentiation between the databases can be seen using MDS.*

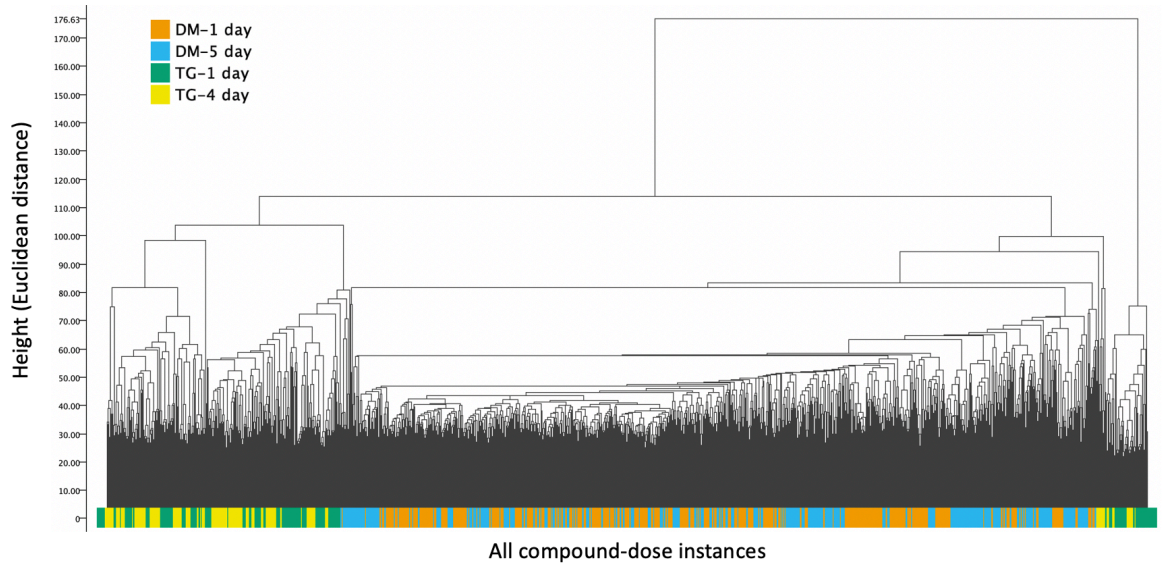



*Figure 2-7 Hierarchical clustering of gene expression data from DrugMatrix and Open TG-GATEs at multiple time points. The two databases are separately clustered. Both databases have trees joining at similar heights (between 30 and 50) showing similar distances between the compound-dose instances. This means that the differences between samples with a database is similar regardless of database, and so the databases can be used to validate hypotheses generated in the other.*

### 2.2.3 Histopathology results

Histopathology observations, as previously noted, form the basis of toxicity and disease phenotypes. Herein, their distribution across organs, databases and scores were displayed, setting the background for any future study based on this work.

For DrugMatrix, Figure 2-8 shows the distributions of distinct histopathology observations across all organs (liver, kidney, mesentery, heart, thigh muscle, spleen and intestine). These reflect all observed phenotypes that occur at least once. The liver and kidney contain the largest number of observations. For Open TG-GATEs, observations were only made of the liver and kidney. It can be seen that Open TG-GATEs has a higher number of histopathology observation, despite the smaller number of samples. Across both, the liver has a higher cumulative number of observations. Therefore, the liver is more likely to be of interest in this work.

Histopathology observations are quantised into ordinal data types, consisting of a rank but no quantitative comparison between each rank. The ranks are not present (no observation was seen), present, minimal, slight, moderate and severe. Figure 2-9 shows the distribution of these scores for both databases. It can be seen that the highest number of observations are not present, representing a large imbalance in class sizes. Both databases tail off at higher severity levels. This creates a problem when modelling and grouping instances together. As such, binarization occurs, usually at the 'minimal' score.[100,102] This cut off is used in further work.

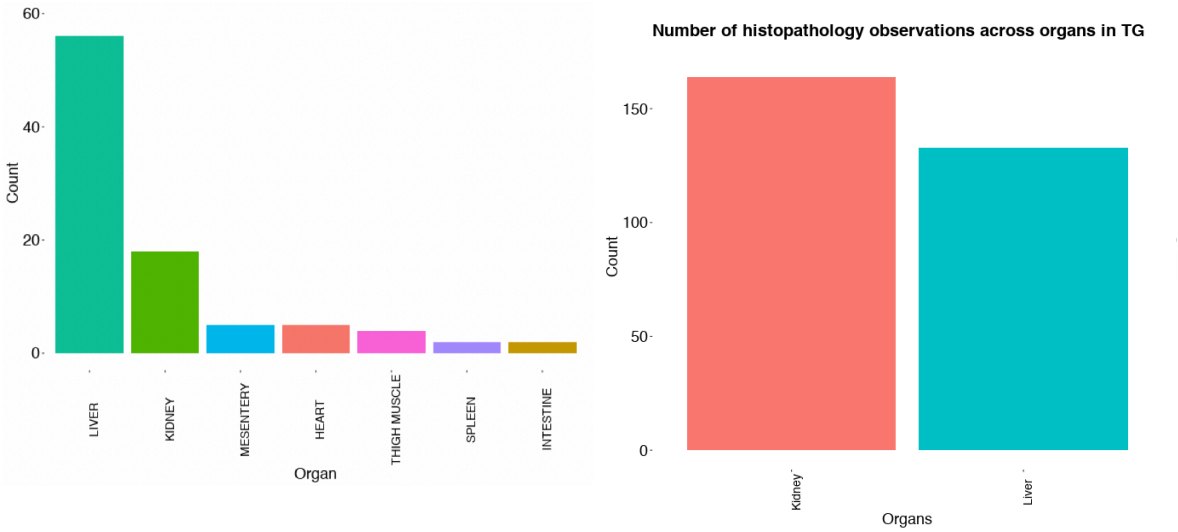



*Figure 2-8 Histopathology observations in DrugMatrix (A) and Open TG-GATEs (B). These show the organs that were viewed microscopically for toxicity phenotypes.*

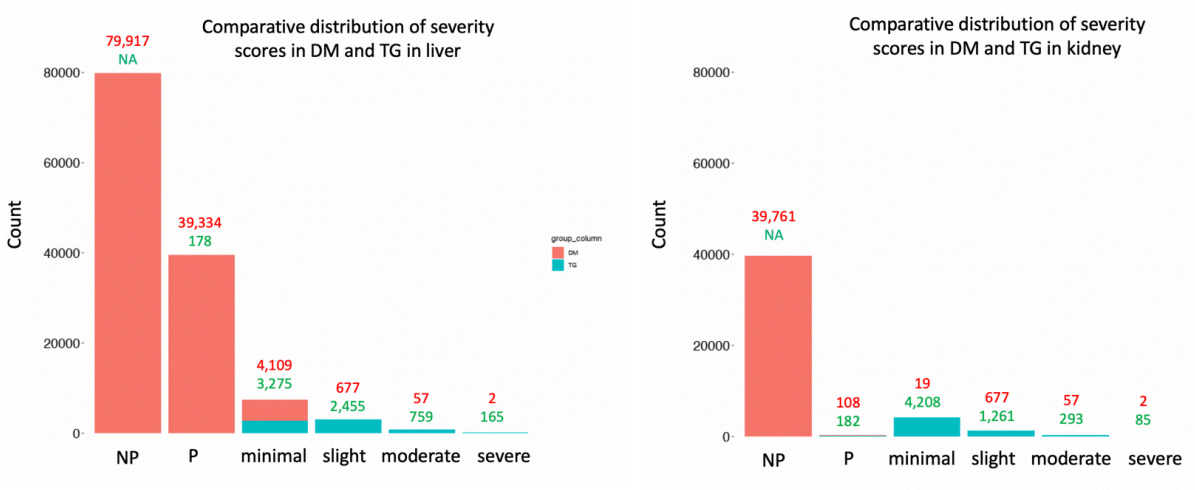



*Figure 2-9 Distribution of severity scores in DrugMatrix (DM) and Open TG-GATEs (TG) for both the liver and kidney. Most observations are negative, and so create a biased dataset.*

In the introduction, it was suggested that modelling single observations is not reflective of the systems-level, *in vivo* response to a stimulus. This hypothesis was tested via determining the independence of each histopathology observation. This was performed using Fisher's Exact Test in a pairwise fashion. The results can be seen in Figure *2-10*. The significant terms shows both inter- and intra- organ histopathology observation dependency in a data-driven manner for all organs. Interestingly, the organ groups are reflected in the significance: observations tend to be dependent on others within the same organ. An interpretation of this is based on cell type. Within an organ, the cell types are the same and so the response is similar on a histopathology observation level. Between tissues, cell types are more different and so the response is different. Inter-organ dependency suggests either failure in experiment (i.e. mixed/contaminated tissues) or a systematic response. This could suggest an inflammatory response that affects the whole organism. Either way, models of histopathology observations should take such relationships into account. Here, this is done using histopathology signatures, created across all observations.

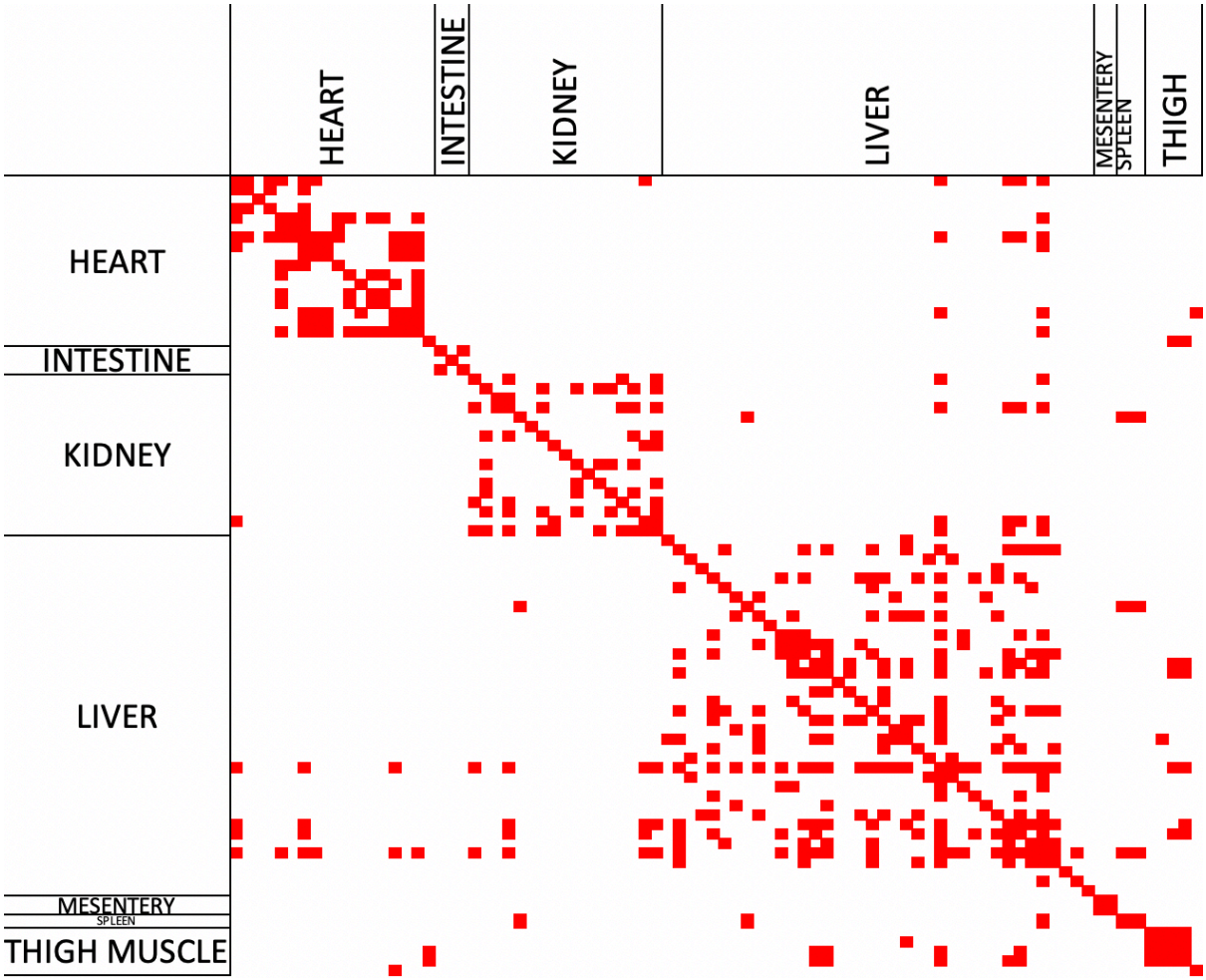



*Figure 2-10 Significance of dependency between histopathology observations. Each histopathology observations in DrugMatrix was tested for dependency against all other, using the Fisher exact test. The FDR corrected p-value was defined to be significant if less than 0.05. Each red square represents a significant p value, meaning the two observations cannot be considered independent. For ease of visualisation, these have been grouped by organ. Inter- and intra- organ dependency can be seen, with higher levels for the latter. Therefore, each observation must be considered with the others when binning to create groups to model, especially those within the same organ.*

In DrugMatrix, 6 groups were found to have a minimum of 5 compound-dose instances, resulting in at least 15 gene expression profiles. The histopathology signatures of these groups are shown in Table 2-4. These 6 groups are non-distinct and share a total of 5 mapped histopathology terms between all of them. All of these are from the liver, which is expected from the distribution of the observations between assays. The compounds comprising this groups are shown in

Table *2-5*. As can be seen, compounds with a wide range of therapeutic purposes are found to cause identical responses, on a histopathology level, in rats.

The appearance of inflammatory cells and infiltration reflects their co-occurrence previously found in this dataset.[102] The groups of compounds associated with different histopathology profiles range in size from 5 to 10 instances and are made up of therapeutically diverse compounds, and there is no link between annotated mode of action class and assigned histopathology group apparent.

The histopathological signatures are defined by only five histopathological observations in total, namely lymphocytic inflammatory cell infiltration, mixed infiltration, lipidic vacuolation (fatty change), glycogen accumulation, and hepatocellular necrosis (Table 2-4), out of 146 possible observations pre-mapping. The many remaining assays are hence too rare, on either an individual or a combined level, to achieve sufficiently large groups for network correlation analysis. Lymphocytic inflammatory cell infiltration is characterised by the recruitment of lymphocytes (white blood cells) by the liver[124] and is associated with the beginning and progression of inflammatory liver diseases.[125] Mixed infiltration, however, covers a wide range of cells that may infiltrate the liver. Both of those processes are commonly occurring phenotypes after exposure to bioactive compounds, especially within the liver where the majority of metabolism occurs.[97,126] As the term "inflammatory" is not present, the cells are non-leukocytes, and so may include other cells such as mast cells.[28] Lipidic vacuolation (fatty change), is a form of steatosis and is characterised by lipids accumulating as vacuoles within cells. [127] Glycogen accumulation is characterised by a distinctive morphologic appearance and it frequently co-occurs with fatty change or cytoplasmic vacuolization.[128] Hepatocellular necrosis, the premature death of liver cells, is a major biomarker of liver injury. It is a frequent finding in chronic and acute liver injury, and if the injury is sustained, leads to fibrosis.[129]

Open TG-GATEs formed 13 groups, with overlapping histopathology observations. These can be seen in Supplementary Table 2 and have a wider selection of histopathology observation. These groups are used in the later chapters of this thesis to probe the biological signal within gene expression data.

*Table 2-4 Histopathology signatures defining 6 toxic groups from DrugMatrix at the 5 day stage Only assays that contain a non-zero score for at least one of the groups are shown, values represent averaged non-binarised values for each group. Each group now will be considered as reflective of their respective combinations of histopathology scores. The terminology of these DrugMatrix observations are in the Histopathology Ontology.*

| Toxic group | Size (number of compound-dose pairs) | Mixed Infiltration | Lipidic vacuolation (fatty change) | Glycogen accumulation | hepatocellular necrosis | lymphocytic inflammatory cell infiltration |
|---|---|---|---|---|---|---|
| Group 1 | 7 | 0.33 | 0.0 | 1.33 | 0.0 | 0.33 |
| Group 2 | 5 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| Group 3 | 10 | 0.33 | 0.0 | 1.33 | 0.0 | 0.83 |
| Group 4 | 6 | 0.66 | 0.33 | 0.17 | 0.0 | 1.0 |
| Group 5 | 10 | 0.83 | 0.0 | 1.0 | 0.33 | 0.83 |
| Group 6 | 5 | 1.0 | 0.0 | 1.33 | 0.67 | 1.0 |

*Table 2-5 Compound-dose instances in each toxic group. Two compounds (imatinib(*) and bithionol(**)) occur in different doses in two groups (Group 2 and 5, and Group 5 and 6, respectively)*

| Group | Compound | DrugMatrix Dose (mg/kg) | class/type |
|---|---|---|---|
| Group 1 | Chlorambucil | 0.6 | chemotherapy, lymphoma |
| | Ethinylestradiol | 10 | estrogen medication |
| | Fluocinolone Acetonide | 2.5 | corticosteroid |
| | Fluvastatin | 5 | statin |
| | Ketoconazole | 114 | antifungal |
| | Mitomycin C | 0.5 | chemotherapy, gastro-intestinal/breast |
| | Spironolactone | 300 | treats fluid build up (heart failure, liver scarring, kidney disease)/steroid |
| Group 2 | Allyl alcohol | 16 | Former herbicide |
| | Imatinib* | 15 | Antineoplastic agent |
| | Megestrol Acetate | 132 | Progestin medication |
| | N-Nitrosodiethylamine | 100 | carcinogen |
| | Pantoprazole | 1100 | Proton pump inhibitor for stomach ulcer treatment |
| Group 3 | Altretamine | 13 | Antineoplastic agent |
| | Clonazepam | 2500 | Treatment for epilepsy and seizures |
| | Clotrimazole | 52 | antifungal |

| | | | |
|---|---|---|---|
| | Cortisone | 206 | Steroid hormone |
| | Diethylstilbestrol | 2.8 | Former estrogen medication |
| | Lomustine | 4.2 | Alkylating antineoplastic agent |
| | Methyl Salicylate | 444 | Analgesic |
| | Mifepristone | 3 | Treatment for Cushing's syndrome and emergency contraception |
| | Nimetazepam | 122 | Hypnotic and anticonvulsant |
| | Salicylic Acid | 223 | Analgesic and anti-inflammitory |
| Group 4 | Artemisinin | 2000 | Treatment for malaria |
| | Bis(2-ethylhexyl)phthalate | 1000 | Medical devices, endocrine disruptor, cardiotoxicity |
| | Carmustine | 4 | Alkylating antineoplastic agent |
| | Raloxifene | 650 | Selective estrogen receptor modulator |
| | Tamoxifen | 2.5 | Selective estrogen receptor modulator |
| | Tosufloxacin | 2000 | Fluoroquinolone antibiotic |
| Group 5 | 17-Methyl testosterone | 2000 | Androgen and anabolic steroid |
| | Acetaminophen | 100 | Analgesic |
| | Balsalazide | 1100 | Anti-inflammatory |
| | Bithionol** | 59 | anthelmintic |
| | Carbamazepine | 490 | Treatment for epilepsy and seizures |
| | Econazole | 43 | antifungal |
| | Ethisterone | 1500 | Progestin medication |
| | Imatinib* | 150 | Antineoplastic agent |
| | Olanzapine | 23 | Antipsychotic medication |
| | Progesterone | 11.3 | Endogenous hormone |
| Group 6 | 1,1-Dichloroethene | 600 | Carcinogen |
| | Atorvastatin | 2.5 | statin |
| | Bithionol** | 333 | anthelmintic |
| | Isoeugenol | 1560 | Sensitiser and allergen |
| | Oxytetracycline | 1500 | antibiotic |

## 2.3 Conclusions

This chapter summarised the chemical, gene expression and histopathology data available within DrugMatrix and Open TG-GATEs, with a view to categorising compound-dose instances for use in modelling and understanding biological mechanisms of action.

Chemical space analysis, when compared to ChEMBL and DrugBank, showed that both DrugMatrix and Open TG-GATEs contain a wide variety of chemical structure and physico-chemical descriptors. It has previously been shown that two highly similar compounds

(Tanimoto similarity > 0.85) only have a one in five change of having similar gene expression profiles.[123] Due to this, the number of compounds and the dissimilar structures within the databases, chemical structure will not be considered in detail for future chapters of this thesis.

Gene expression data analysis showed the high amount of data that is internally consistent (within a database). Therefore each database can be used separately, as in the next two chapters of this thesis. Gene expression data is used from one database to determine histopathology – gene expression associations. These associations can then be tested in the other database for consistency/validation.

Analysis of histopathology observations showed the largest occurrence of observations to be in the liver, Open TG-GATEs has wider ranging histopathologies, and that the observations are dependent on each other. As such, the next chapters of this work use the histopathology signatures and groups of compound-dose instances to ensure that a systems-level approach is considered. The DrugMatrix database is used to generate histopathology – gene associations and these are tested in Open TG-GATEs, as the range of histopathologies from DrugMatrix is captured in this.

# 3. Consistency evaluation of mechanistic hypotheses for rat liver histopathology readouts from DrugMatrix and Open TG-GATEs considering co-dependency of observations

This work considers co-occurring histopathology observations as signatures and aims to determine novel and known associations between the transcriptomics and histopathology level. Associations were furthermore derived from the DrugMatrix database and validated against Open TG-GATEs and histopathology signatures for conserved associations compared. Hence, for the domain of rat *in vivo* liver injury, we aimed to investigate the extent to which complex histopathology signatures are conserved between both databases (and hence more generally for such data from different sources), what those conserved signatures are, and how they are mechanistically driven on the gene expression level. This is based on using data driven approaches to compliment known biological associations.

Initial work based on DrugMatrix and Open TG-GATEs included the determination of predictive models for carcinogens. Using labels from 2 year rat carcinogenicity studies, a 37 gene-set signature was determined from gene expression data taken at day 5 with daily exposure.[130] Whilst carcinogenicity is outside the focus of this work, the high performance of the models (sensitivity and specificity of 86% and 81% respectively on a validation set) shows that there is a signal from microarrays readouts that translates to *in vivo* toxicity. On the mechanistic level, only 27 of the 37 signature genes were annotated in Gene Ontology, which represent various aspects of antiapoptotic effects and proliferation. However, when the gene signatures were tested on non-DrugMatrix data, it was found that "the decreased level of test sensitivity and specificity are not likely sufficient to justify routine use of the signatures as tested outside their laboratory of origin".[131] This finding makes the importance of generating consistent data, with enough coverage, clear to be able to arrive at truly predictive toxicity models. Subsequently, using Principle Component Analysis and prediction of mode-of-action on gene expression data (cytotoxicant, enzyme inducer, PPARα agonist or "other"), Kanki *et al.* determined a 106 gene-set that accurately classified carcinogens with respect to time, dose and mode-of-action.[132] Biological interpretation of this gene set *via* Ingenuity Pathway Analysis elucidated known associations to cancers, such as the geneset's connection to regulated gene markers (MAPKs and PI3/AKT signalling) with downstream perturbation of hepatic system development and function, lipid metabolism, and organ morphology.[65] This model was not externally validated, meaning its prospective predictive power remains unclear.

Both DrugMatrix and Open TG-GATEs data have also been used in the non-cancer field to model and predict specific organ histopathology readouts. This included modelling presence or absence of histopathology observations, their severity, and using ensemble models of the observations to predict a general "toxic" label.[133,134] In one study, the severity of four different histopathology observations (necrosis, hypertrophy, cell infiltration and leukocytic change) was determined using the Random Forest algorithm, based on genes selected *via* sparse Linear Discriminant Analysis.[133] Subsequently, the biological relevance of the selected genes was determined *via* enrichment of Gene Ontology (GO) terms, which were taken from the top level of GO and not mechanism focused (e.g. "response to xenobiotic stimulus"). In a separate study, an ensemble model was built on 21 histopathology endpoints from Open TG-GATEs, employing a k-nearest neighbour approach based on compound-induced gene expression data and resulting in an AUC of 88% (sensitivity 0.84, specificity 0.83).[134] Whilst these approaches produced relatively accurate models, they do not provide in-depth biological interpretations for the histopathology observations. This study aims to explain the associations between gene expression data and histopathologies.

The workflow of the current work is shown in Figure 3-1. Histopathology phenotypes from DrugMatrix and Open TG-GATEs were expressed as *signatures* to define toxic sets (a). In order to obtain unique signatures with a sufficient number of associated data points, histopathology observations from DrugMatrix were mapped to the HPATH histopathology ontology (b).[27] This ontology, created by the eTox consortium, classifies and categorises histopathology observations and morphologies from animal studies. The created signatures defined toxic groups of compounds, whose differentially expressed genes after treatment (c), and resulting enriched pathways, and pharmacological profile were determined (d). Co-expression networks were created for each toxic group using the WGCNA approach discussed above, which revealed known and novel associations that were consistent between DrugMatrix and Open TG-GATEs databases (e). These associations were determined on a gene, pathway, GO term, disease and transcription factor level, and hence to the best of knowledge of the author this represents the first study that establishes consistency between the two major toxicogenomics databases available on a large scale for liver toxicity readouts, while also considering multi-endpoint histopathology definitions. Throughout this work, a histopathology signature is defined as the group of observed histopathologies whereas the toxic group defines the compound-dose instances and their gene expression that make up each histopathology signature.

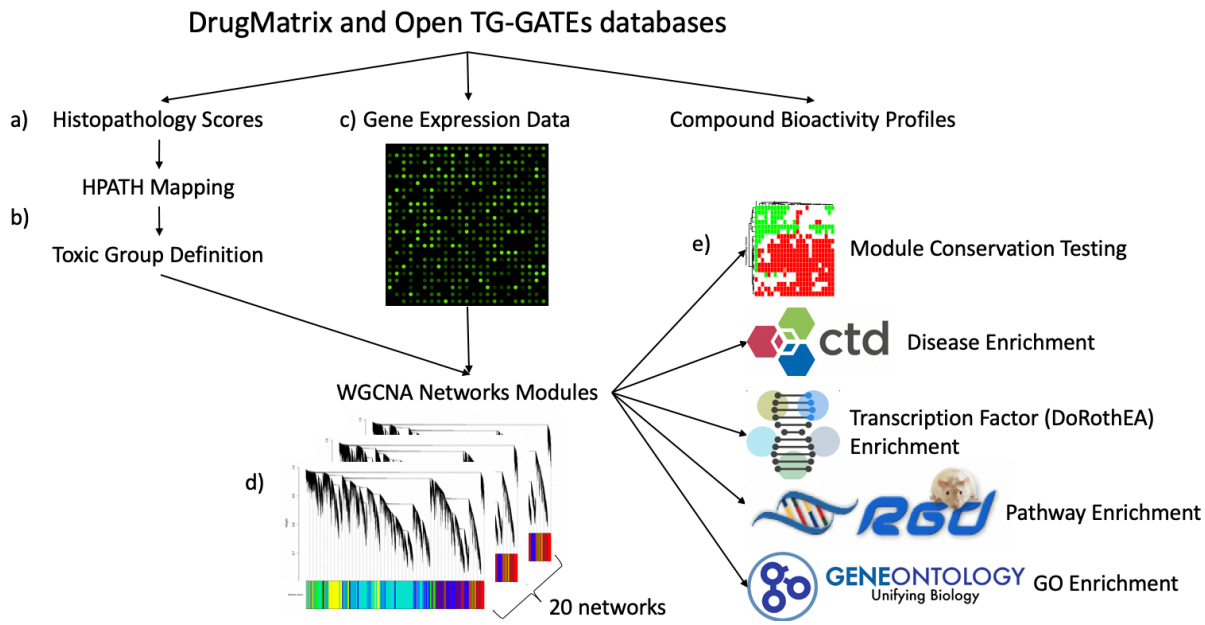



*Figure 3-1 Workflow of this study. Compound-dose groups are segregated by their histopathology observation and the resultant gene-histopathology associations are tested for conservation between databases and biological meaning. Histopathology observations (a) are mapped to HPATH and define toxic groups (b). Differentially expressed genes and compound bioactivity profiles are determined (c) and co-expression networks created (d). The resultant networks are tested between databases (e) and their biological meaning determined by enrichment testing from gene-disease (the Comparative Toxicogenomics Database), gene-transcription factor (DoRothEA), gene-pathway (Rat Genome Database), and gene-GO term (Gene Ontology) data.*

An important confirmation of gene-gene associations is whether their behaviours are mirrored at the protein level. To this end, module (groups of genes) associations can be compared to known gene products (protein) interactions. Numerous databases of protein-protein interactions (PPIs) are within the public domain.[59] Within this study, it is key to use data from rat models, and as such, the String database was used.[178]

## 3.1 Methods

The histopathology signatures and toxic groups were used as generated in Chapter 2.

### 3.1.1 Bioactivity

The matrix of 79 bioactivities for all compounds was downloaded from the DrugMatrix database.[120] This matrix contains activity (or inactivity) values for all the compounds. The

matrix was clustered using Euclidean distances and complete linkage in R (version 3.5) with package 'ComplexHeatmap'.[135]

### 3.1.2 Weighted Gene Coexpression Network Analysis (WGCNA)

Coexpression network analysis was performed using the WGCNA package (version 1.66) in R (version 3.5.1). All gene expression CEL files from each histopathology signature were used to generate a network, (Groups 1-6 defined in Table 2-4 and controls for DrugMatrix, and the Open TG-GATEs groups). Raw RMA-adjusted gene expression values (as previously generated in Chapter 2) were used to generate the correlation matrix. The softpower was determined by comparing the scale-free topology and mean connectivity and was subsequently set to be 10 in each case. Unsigned networks were created for all groups (as default). Modules were formed using the FlastClust function with average linkage. The tree was cut at height = 0.99 for tighter module formation, and the remaining parameters were used as their default values.

Module conservation was determined by permutation testing of mean densities and correlations, both inter- and intra- modular, of connectivity, adjacency matrices and eigengene values.[83] 30 permutations were performed, with a maximum module size of 1,000 genes. Zsummary scores are used as a proxy for conservation. Zsummary scores greater than ten were taken to show significant conservation, and those less than 2 to show no evidence of conservation.[83] A module of randomly selected genes (n=1,000) was created 30 times and the same permutation test was applied to determine the false discovery rate of conservation scores. This was performed in line with the WGCNA protocol to create and test coexpression network modules.

### 3.1.3 Pathway enrichment

Gene Ontology biological processes and pathway ontologies were downloaded from the Rat Genome Database.[62] Differentially expressed genes and genes in modules were tested for enrichment using the 'Enrichr' package.[21] Significance of enrichment was determined using Fisher's exact test as implemented in Enrichr, with all measured genes (from the original microarray) used as a background. A Benjamini-Hochberg multiple testing correction was applied in R and an adjusted p-value cut-off of 0.05 defined significance. Significant GO terms were visualised with RamiGO.[136] Pathway ontologies were visualised with OntologyX.[137]

### 3.1.4 Disease enrichment

All disease-gene associations were retrieved from the Chemical Toxicogenomics Database (CTD)[48] and the distribution of the inference scores of these disease-gene associations is shown in Figure 3-2. As can be seen, the distribution has a long tail, meaning that very few disease-gene inferences have scores above 200. Three cut-offs of the inference score (50, 100, and the top 5%) were used to decide whether genes and diseases were associated. Gene sets for each disease instance were created and filtered to remove those with fewer than 10 genes. Disease enrichment for each module was performed using the same statistical approach as described above for the pathway enrichment. Diseases-module associations were considered significant if they were present at both the 100 and 5% inference score cut-offs.

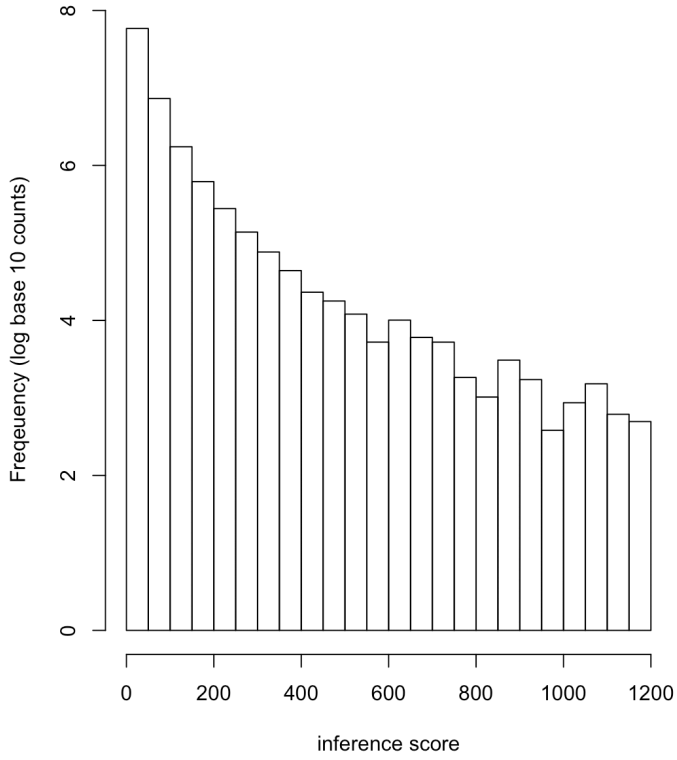



*Figure 3-2 Distribution of interference score of gene-disease association from the comparative toxicogenomics database. The y axis is a logarithmic scale (base 10) and so there is a long tail.*

### 3.1.5 Transcription factor enrichment

Transcription-gene associations were downloaded from DoRothEA[138] using the confidence limits of A-C, as recommended to obtain higher confidence associations.[139] Gene

sets for each transcription factor were created, filtered and tested for significant enrichment using the same statistical approach as in the disease enrichment. This step was performed mainly to determine module specificity – genes form modules if they are co-expressed and therefore are likely to share transcription factors. The incomplete nature of the database allows for positive enrichments to be found, though without being able to draw conclusions from the absence of enrichment.

#### 3.1.6 Protein – Protein interaction significance

Protein – protein interactions were downloaded from the String database.[178] Specifically, 22,763 interactions that occur within rats were downloaded. Random subsets of proteins were selected, with sample sizes the same those in the defined modules. Each sample was permuted to determine all pairwise interactions. The number of known interactions (as defined by the String database) was compared to the number of true interactions from the each module. The random sub-setting was repeated 500 times to determine significance, against each module from each toxic group. A KS test was performed in R, to determine whether the edges in the module subgraphs (between genes) were reflected in the biological interactions (PPI network).

## 3.2 Results

### 3.2.1 Using Bioactivities, Differentially Expressed Genes and Pathway Enrichment to Discriminate between Histopathological Signatures

We firstly analysed annotated protein activities, differentially expressed genes, and pathway enrichments for their ability to discriminate between toxic compound groups.

We found when clustering based on annotated bioactivities that this did not discriminate between the toxic groups (Figure 3-3), where members of different toxic groups can be found spread out widely in bioactivity space. Hence, protein target activity alone is an insufficient predictor of particular histopathology readout profiles.

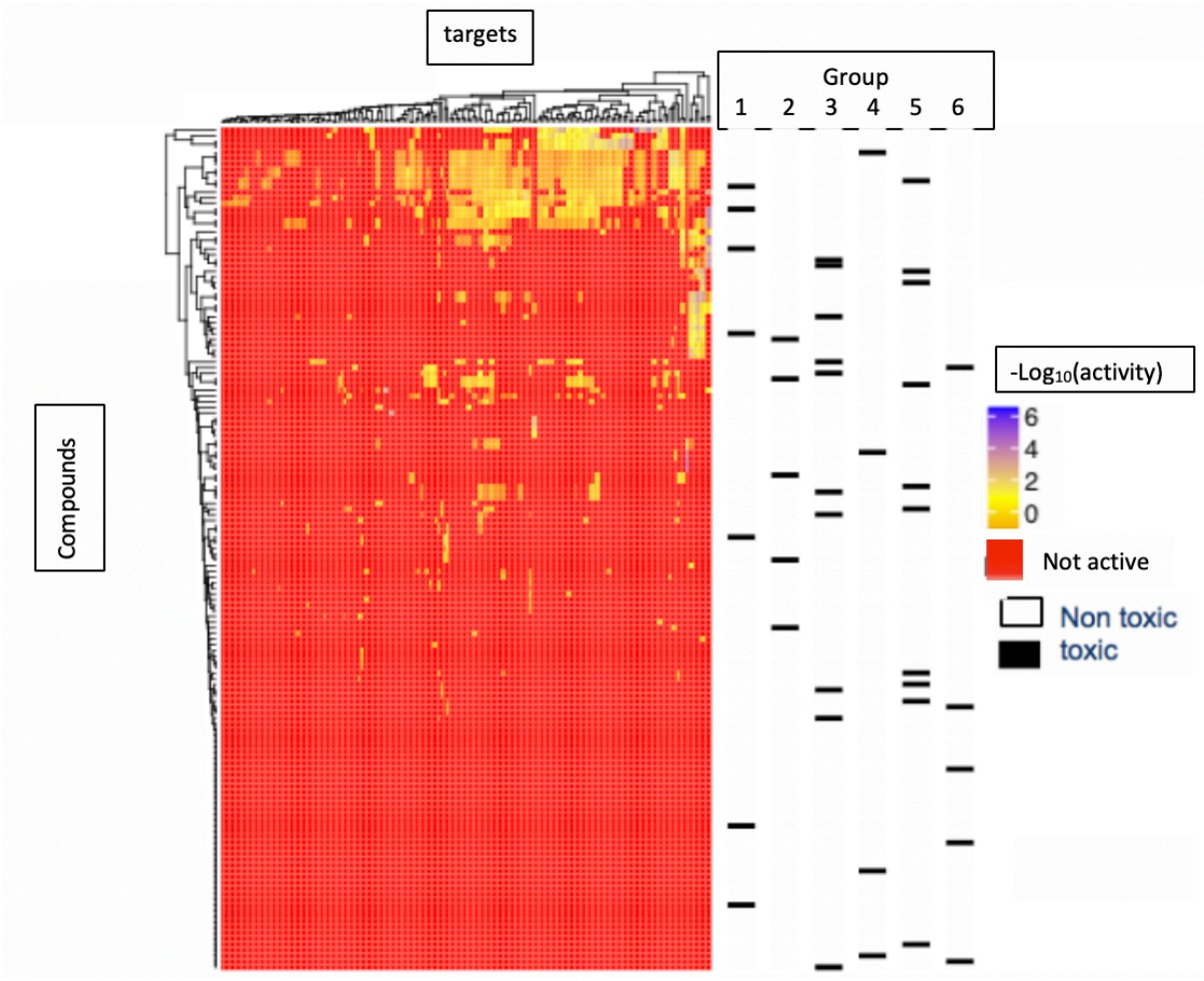



*Figure 3-3 the bioactivity profiles of the compounds across 79 assays. It fails to discriminate between the toxic and non toxic instances. The red colour represents no activity and was defined in DrugMatrix.*

Similarly, differentially expressed genes by themselves failed to group between the different toxic compound classes with each other, and separately from non-toxic instances (Figure 3-4). It can be observed that individual genes were in most cases *either* up- *or* downregulated across compound-dose instances, but not upregulated in some compound-dose instances, and downregulated in others. This finding has previously been observed by Gusenleitner *et al.*[140] when predicting the long term carcinogenicity of compounds using the DrugMatrix database. Their analysis noted that this effect was strongest when in response to carcinogens. Whilst this behaviour can be rationalized for genes that have low expression levels (reaching the detection limit of the microarray in case of no compound treatment and upregulation after treatment), there is no obvious explanation for commonly expressed genes.

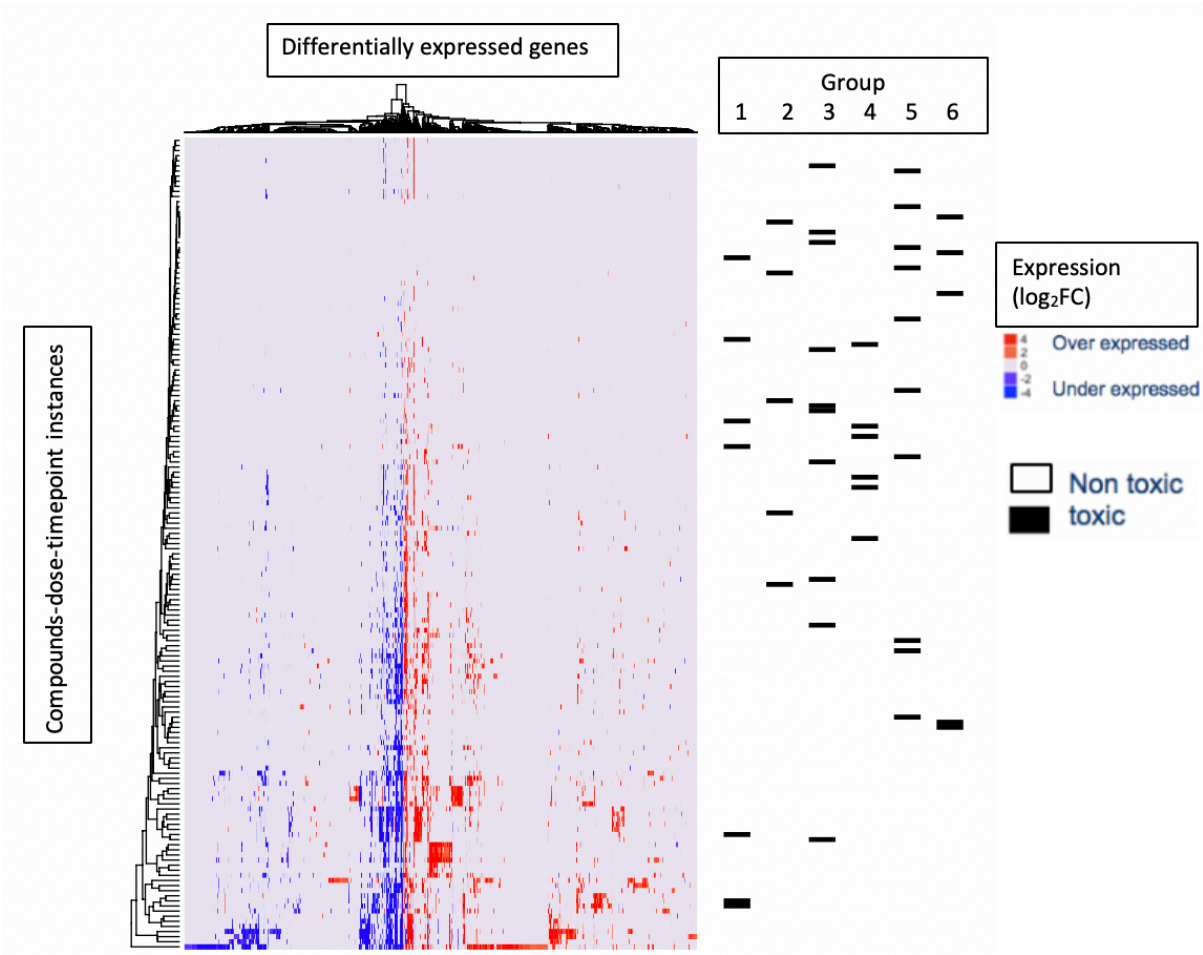



*Figure 3-4 the differentially expressed genes for each compound-dose instance at the five-day stage. As the black bar on the righthand side shows, there is very little differentiation between toxic and non-toxic instances. The enrichment of these genes can be seen in the supplementary information (Supplementary Table 3). Similarly to the clustering in bioactivity space, gene expression does not clarify the distinction between the toxic and non-toxic sets. Note, genes that were not differentially expressed in any compound-dose-time point instance and those with zero variance were filtered out for ease of representation*

The differentially expressed genes per toxic group were next tested for GO pathway enrichment. This resulted in 166 enriched pathways over each group (see Supplementary Table 3). These pathways are compound-specific and their applicability to each group is not clear. The large number of enriched pathways prevent deconvolution of the signal.

Hence we can conclude that neither protein activities, not clustering based on gene expression, or pathway enrichments were able to provide mechanistic hypotheses for the groups of toxic compounds defined by histopathology profiles in this work.

### 3.2.2 WGCNA analysis

Co-expression networks were then made to investigate how gene-gene relationships vary within and between different toxic groups. This created module-histopathology associations on the gene and pathways level.

We generated co-expression networks for each of the six toxic compound groups compared to control. Each network forms hierarchical clusters of genes, which are highly co-expressed, and are named after colours for ease of reference. The modules defined in each network were then tested against each other (all against all) and against the Open TG-GATEs groups to determine their conservation (i.e. if the module genes have similar behaviour compared to each other and all other measured genes, in different networks). These results of which are shown in Figure 3-5 for compound group 1, and Supplementary Figure 2 for groups 2-6, respectively. High conservation (Zsummary score) indicates that inter- and intra-modular network characteristics are similar in different networks. Complete segregation between DrugMatrix and Open TG-GATEs networks can be seen. This is to be expected due to the differences (in time, location and exact experimental protocol) of the experiments. Similarly, there were more conserved modules within all six of the DrugMatrix groups. An average of 49% of modules were conserved with DrugMatrix compared to 6% with Open TG-GATEs. Additionally, there were more histopathology endpoints in Open TG-GATEs (13 compared to 6) which cover a larger range of observed histopathology. Hence, we can conclude that the two databases are distinct but there is shared gene expression signal between them (conserved modules).

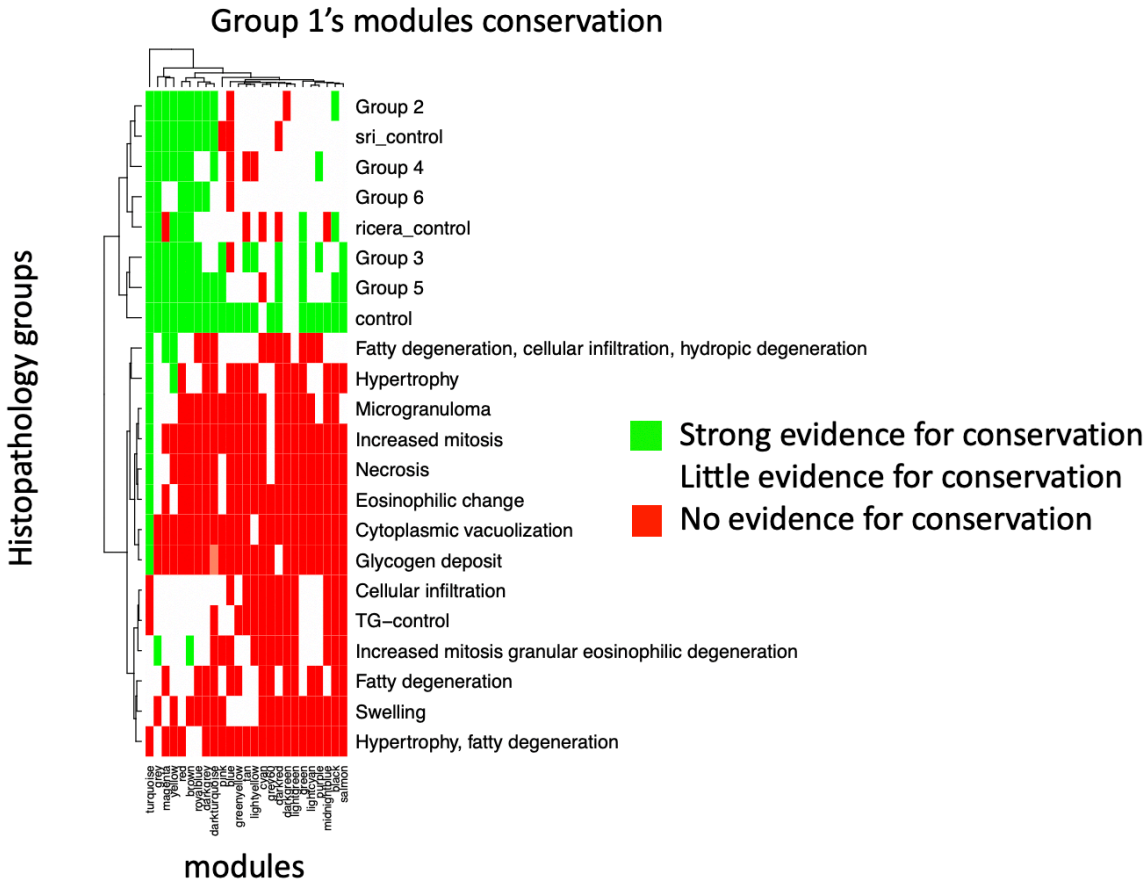



*Figure 3-5 Results of permutation testing of modules for Group 1, representing which modules defined by Group 1 are conserved with different histopathology groups in both DrugMatrix and Open TG-GATES. The separation of DrugMatrix and Open TG-GATES is apparent, and the conservation of particular modules between databases shown. The Open TG-GATEs group representing "fatty degeneration, cellular infiltration, hydropic degeneration" shows the highest inter-database concordance. SRI and Ricera represent different external controls that were included in the DrugMatrix database.*

#### 3.2.2.1 Conservation of compound group 1, represented by mixed infiltration, glycogen accumulation and lymphocytic inflammatory cell infiltration

We next analysed compound group 1 in detail to determine and understand the associations between group 1, the controls, and the other toxic groups. These associations are on the gene, pathway and histopathology level.

The histopathology signature of Group 1 is defined by the original DM observations of mixed infiltration (average score of 0.33), glycogen accumulation (average score of 1.33) and

lymphocytic inflammatory cell infiltration (average score of 0.33). Its coexpression network formed 25 modules in total.

*Comparison of generated modules of Group 1 with modules of the control group*

First, we tested the modules defined by group one to see what was not conserved with the controls (no compound perturbation). These modules were tested for their biological meaning, disease association and transcription factor enrichment.

Table 3-1 shows all modules that are either not-conserved with control (i.e. represent perturbations from the non-compound-induced state) and the modules that are conserved in networks from Open TG-GATEs (i.e. genes act the same way in relation to each other in the external database). Three modules are not conserved between Group 1 and the control: lightgreen, darkgreen, and cyan. Lightgreen had no significantly enriched pathways. However, there were 11 significantly enriched diseases for this module (shown in Table 3-2) seven of which are directly related to the definition of Group 1: ‘chemical and drug-induced liver injury’, ‘hepatomegaly’, ‘liver diseases’, ‘fatty liver’, ‘inflammation’, ‘necrosis’, and ‘hypertrophy’. Hence, we can conclude that disease enrichment can be a useful analysis method to associate modules to general toxicity terms.

*Table 3-1 Conserved modules between Group 1 (from DrugMatrix), control group and Open TG-GATEs histopathology groups. High Zsummary scores (>10) indicate module conservation between databases, while low scores indicate no conservation and so modules not conserved with the control are shown. Each Open TG-GATEs group is annotated with its histopathology signature. There are 3 modules not conserved with control (cyan, darkgreen and lightgreen) and 4 modules that are conserved between Group 1 and Open TG-GATEs groups, to varying degrees. The signature that is conserved in the highest number of modules is 'fatty degeneration, cellular infiltration and hydropic degeneration'. This shows clear parallels with histopathology signature of group 1, and hence we can conclude that coexpression network module comparison shows concordance of gene expression and histopathology signature.*

| Module | Zsummary score | Histopathology signature from Open TG-GATEs group |
|---|---|---|
| cyan | 8.02 | Not-conserved with DrugMatrix Control |
| darkgreen | 7.10 | Not-conserved with DrugMatrix Control |
| lightgreen | 8.51 | Not-conserved with DrugMatrix Control |
| brown | 10.90 | Increased mitosis, and granular eosinophilic Degeneration |
| magenta | 16.00 | Fatty Degeneration, Cellular infiltration, and hydropic Degeneration |
| turquoise | 26.83 | Microgranuloma |
| | 35.11 | Hypertrophy |
| | 12.14 | Cytoplasmic Vacuolization |
| | 13.29 | Fatty Degeneration, Cellular infiltration, and hydropic Degeneration |
| | 21.61 | Necrosis |
| | 20.30 | Eosinophilic Change |
| | 11.78 | Glycogen deposit |
| | 28.28 | Increased mitosis |
| yellow | 10.72 | Hypertrophy |
| | 16.34 | Fatty Degeneration, Cellular infiltration, and hydropic Degeneration |

*Table 3-2 Disease enrichment of the lightgreen module, seven of which (chemical and drug induced liver injury, hepatomegaly, necrosis, fatty liver, hyperplasia, liver diseases and hypertrophy) are directly related to group 1's histopathology signature.*

| Disease term | Number of genes in module | Number of genes in pathway | p value | Adjusted p value |
|---|---|---|---|---|
| Chemical and Drug Induced Liver Injury | 35 | 3645 | $3.80 \times 10^{-8}$ | $9.25 \times 10^{-6}$ |
| Hepatomegaly | 21 | 1366 | $5.41 \times 10^{-8}$ | $9.25 \times 10^{-6}$ |
| Inflammation | 27 | 2443 | $3.07 \times 10^{-7}$ | $3.49 \times 10^{-5}$ |
| Necrosis | 35 | 4048 | $6.41 \times 10^{-7}$ | $5.48 \times 10^{-5}$ |
| Weight Loss | 24 | 2525 | $3.07 \times 10^{-5}$ | $2.10 \times 10^{-3}$ |
| Kidney Diseases | 17 | 1410 | $3.94 \times 10^{-5}$ | $2.24 \times 10^{-3}$ |
| Fatty Liver | 14 | 1009 | $5.06 \times 10^{-5}$ | $2.47 \times 10^{-3}$ |
| Hyperplasia | 20 | 1938 | $6.22 \times 10^{-5}$ | $2.66 \times 10^{-3}$ |
| Prenatal Exposure Delayed Effects | 24 | 2675 | $8.12 \times 10^{-5}$ | $3.08 \times 10^{-3}$ |
| Liver Diseases | 11 | 814 | $4.68 \times 10^{-4}$ | $1.60 \times 10^{-2}$ |
| Hypertrophy | 9 | 631 | $1.12 \times 10^{-3}$ | $3.48 \times 10^{-2}$ |

We furthermore analysed hub genes (genes with highest degree of connectivity) for each module, i.e. those genes which are thought to be influential drivers of the function of one part of a network. Figure 3-6 shows the hub genes for the lightgreen module, which are associated with lipid binding, tubulin binding, proteolysis and ion transport. Despite no pathways being significantly enriched, its hub genes hence show a connection to the toxic definition in Group 1, also showing that network analysis can provide insight into gene modules beyond pathway enrichment alone.[58]

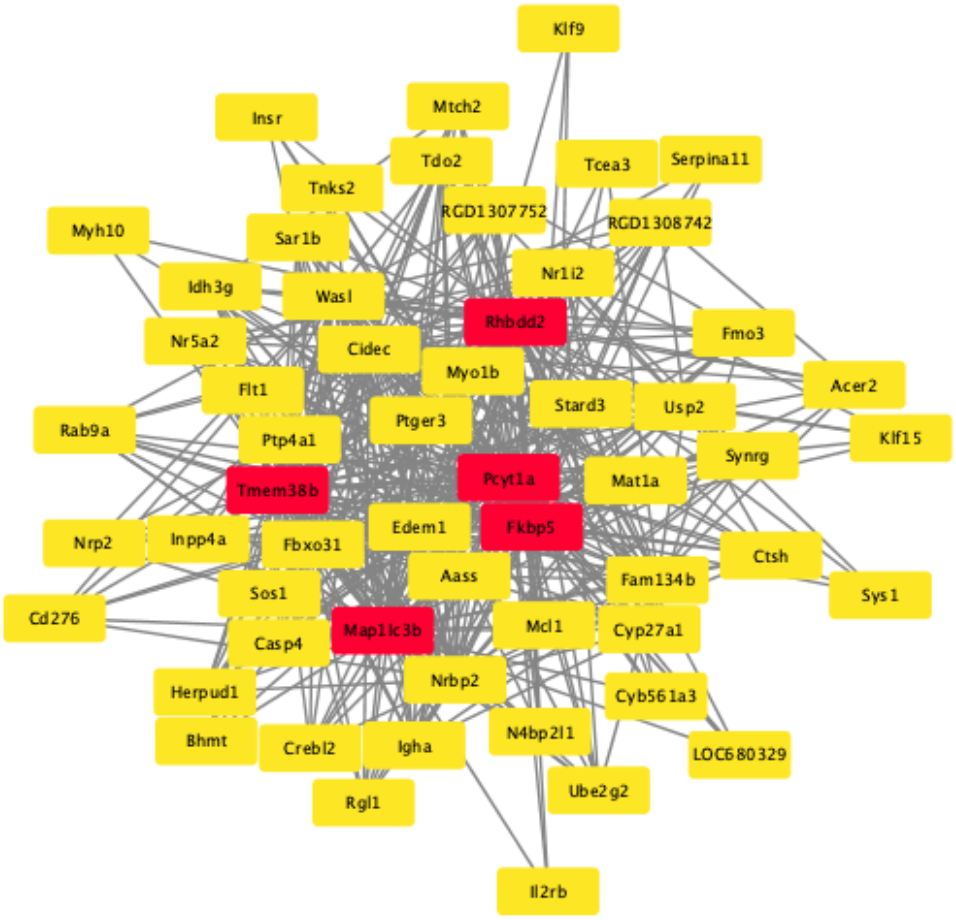



*Figure 3-6 Lightgreen module for the Group 1 case study, based on visualising the sub-graph of the coexpression network (to show the gene in the module). The five genes with highest degree ('hub genes', displayed in red) are associated with lipid binding, tubulin binding, proteolysis and ion transport.*

The darkgreen module has five significantly enriched pathways, as shown in Figure 3-7, namely the 'fructose and mannose', 'pentose phosphate', and 'gluconeogenesis' metabolic pathways, and the 'hereditary fructose intolerance syndrome' and 'fructosuria' disease pathways. Of these, there is literature evidence to support that the pentose phosphate and gluconeogenesis pathways have roles in chemical- and drug- induced liver injury, and promoting hepatic steatosis (fatty liver).[141] Specifically, the pentose phosphate pathway parallels lipogenesis since it is a major source of NADPH (and pentoses) and NADPH is consumed for reductive biosynthesis.[142] The importance of fructose metabolism is a known cause of liver diseases.[143] However, the specific pathway of 'fructose and mannose metabolic pathway' may play a more significant role in the manifestation of the phenotype. The enriched disease pathways are fructose intolerance and fructosuria. The significance of fructose in hepatotoxicity is its alteration of the concentration of metabolites which promotes

intrahepatic fat deposition and hyperuricemia.[144,145] These two pathways help to provide a mechanistic background with the observed glycogen accumulation which was part of the definition of compound Group 1.

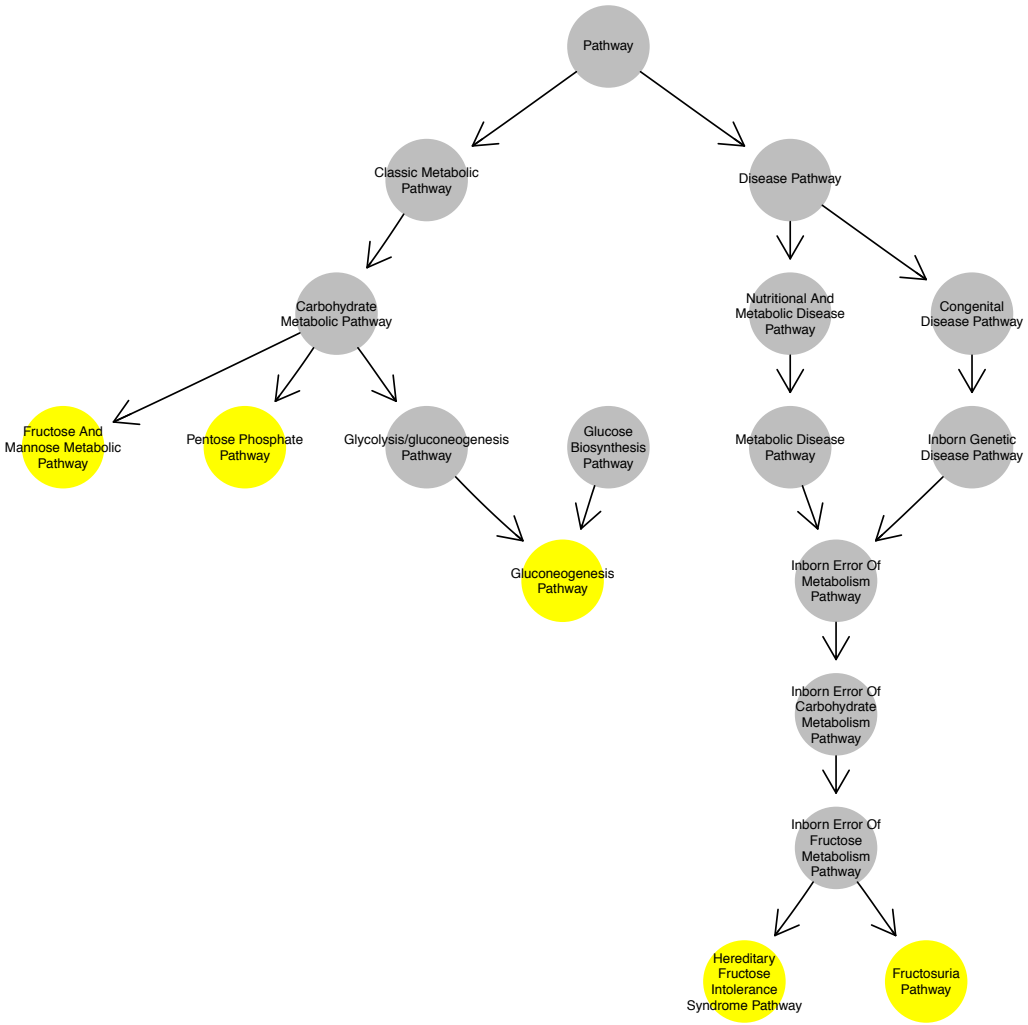



*Figure 3-7 Ontology of enriched pathways in the darkgreen module of compound Group 1, which was conserved within the compound group, but different from control. The Pentose phosphate pathway is linked to 'chemical and drug induced liver injury' via the four genes ALDOB, PGM1, RGN, and TALDO1. Gluconeogenesis has been shown, via mouse knock out studies, to promote hepatic steatosis (fatty liver). Fructose can enter liver hepatic cells without the prescence of insulin (the rate limiting step for glucose). This provokes the change in concentration of several metabolites, ultimately being linked to intrahepatic fat deposition and hyperuricemia.[144,145] Given that the histopathology readout of compound Group 1 is partly defined by glycogen accumulation, this module provides a mechanistic hypothesis of steps leading to this effect.*

The two remaining histopathology observations that define the histopathology signature of Group 1 are ‘mixed cell cellular infiltration’ and ‘lymphocytic inflammatory cell infiltration’. We have next analysed the cyan module in order to explain this, as shown in Table 3-3, which shows the Gene Ontology enrichment for the cyan module. The one enriched pathway is “cellular response to corticotropin-releasing hormone stimulus”. Corticotropin-releasing hormone is a peptide hormone involved in stress response. This, alongside the more generic ‘inflammatory response’ GO term, demonstrates the resulting stress response to the compound exposure. There are 180 unique and significant disease associations for this module, reflecting the generic stress response from a system when changing to a disease state. Additionally, there are 7 enriched transcription factors, namely GATA2, RELA, NFKB1, STAT3, ATF2, STAT1, and CREB1, in the cyan modules. This supports the biological interpretation of the data driven approach – namely that there is evidence that specific transcription factors modulate genes that form part of particular modules. Additionally, NFKB1 is central to the role of liver homeostasis, and regulation of inflammation, fibrosis and carcinogenesis.[146] Hence, the enriched pathways from these modules show biological signal for the observed histopathology. Similar to its enriched pathways, the cyan module’s hub genes are highly connected to stress response via interleukin-1 and 10 (Supplementary Table 4), showing that the systems stress response has been activated. The top hub gene, Cxcl2 is involved in ‘response to lipopolysaccharide’. This is highly associated with chemical- and drug- induced liver injury, and inflammation (over 250 references in the Chemical Toxicogenomic Database).[48]

*Table 3-3 All enriched GO terms in the cyan module of Group 1, which not conserved with the control. Corticotropin-releasing hormone is a peptide hormone involved in stress response, as is that of inflammatory response. This parallels the stress response seen in Group 1's definition (inflammatory cell infiltration). TermID is the GO term ID and the multiple hypothesis adjustment is Benjamini-Hochberg.*

| TermID | Number of genes in module and pathway | Total number of genes in pathway | p value | Adjusted p value |
|---|---|---|---|---|
| GO:0006954 inflammatory response | 10 | 215 | $3.77 \times 10^{-7}$ | 0.005 |
| GO:0071376 cellular response to corticotropin-releasing hormone stimulus | 3 | 5 | $1.78 \times 10^{-6}$ | 0.011 |
| GO:0070098 chemokine-mediated signalling pathway | 5 | 46 | $6.11 \times 10^{-6}$ | 0.025 |

In this section we found three modules for the first histopathological signature (Group 1) which were not conserved with the control group and which hence would form hypotheses for the mechanistic basis of the observed toxicity signature. Whilst the histopathology observations have been discussed individually, they are only formed when generating a group of compounds with identical signatures. These three modules reveal known and novel associations, namely the specific involvement of fructose and mannose metabolic pathways and a systemic stress reaction.

However, the robustness of these associations next needs to be confirmed in comparison to other databases to confirm the consistency of modules derived for compound Group 1 with external histopathology observations, to show the general applicability and validate the prospective value of the approach.

*Comparison of modules for Group 1 with groups from Open TG-GATEs*

We next analysed the conservations of Group 1's modules with those from Open TG-GATEs in order to determine whether the histopathology observations are conserved on a gene level

between databases. The results are shown in Figure 3-8. It can be seen that each toxic group is conserved to differing extents with those from Open TG-GATEs. Most commonly, 'fatty degeneration, cellular infiltration and hydropic degeneration' is conserved with DrugMatrix groups, which broadly match their definition.

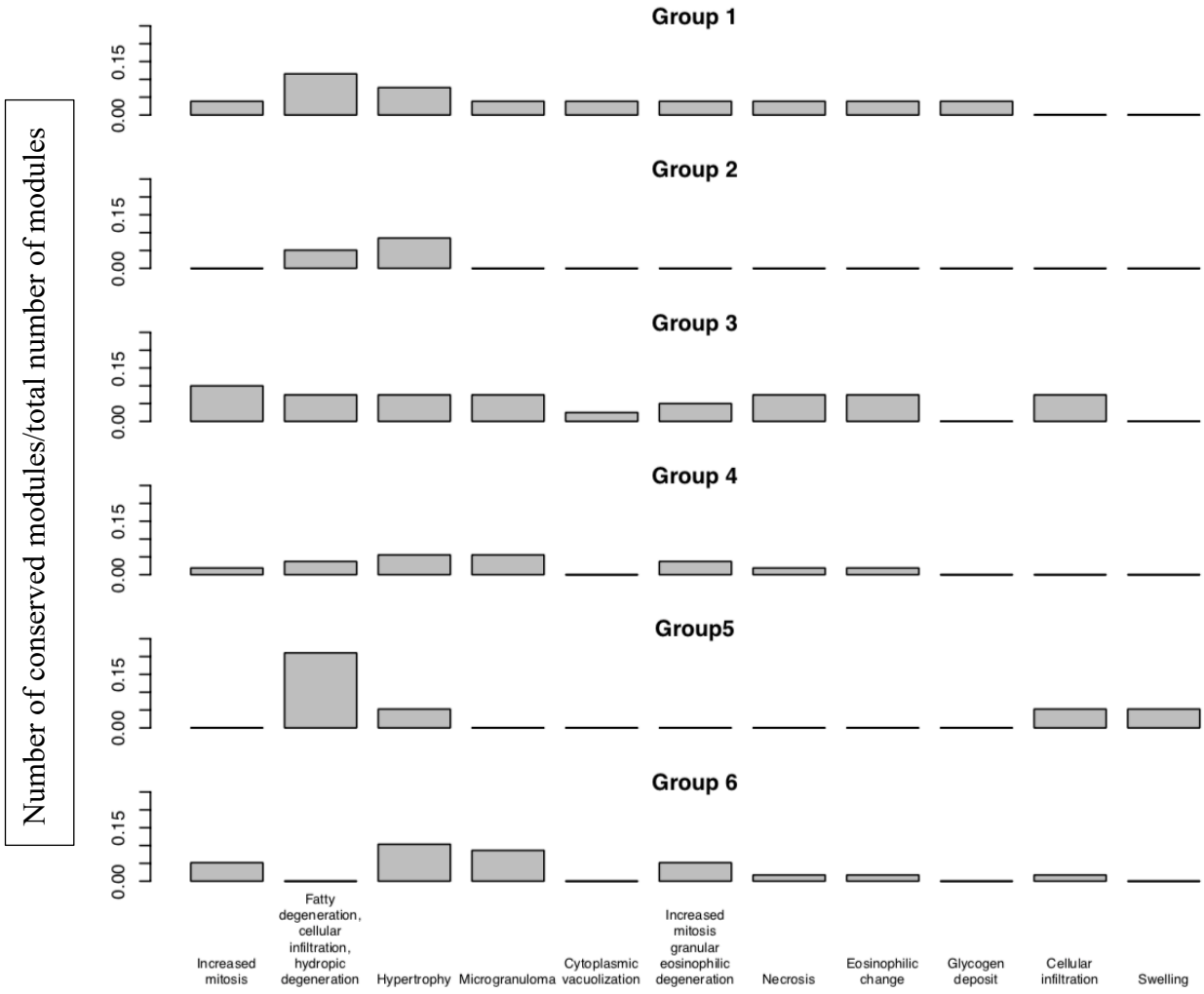



*Figure 3-8 Distribution of conserved modules across Open TG-GATEs histopathology signatures for Groups 1-6. 'hypertrophy' and 'fatty degeneration, cellular infiltration, and hydropic degeneration' both consistently show the highest amount of conservation, reflecting concordance of histpathology via co-expression networks between database*

The cyan, lightgreen and darkgreen modules are not externally confirmed with other toxic groups (Figure 3-5), even though these have overlapping toxic definitions (Table 2-4). Whilst these are the only modules not conserved with the control group, the inter-database conservations are still important, as these reflect the behaviour of the genes, not their absolute expression.

Histopathology-specific gene expression network concordance between database was determined, through conserved modules. These modules are, as shown in Table 3-1, namely the brown, yellow, magenta and turquoise modules. Figure 3-9 show their enriched pathways and GO terms. As is clearly visible, there is a wide range of functional biology within these modules, which we assume to represent the underlying/background biology present within similar toxic groups (and, indeed, between databases).

Conservation between databases is quantified as the fraction of modules that are conserved by histopathology signatures from different databases, and the signatures are compared. Figure 3-8 shows this frequency of conserved modules between Group 1 and those from Open TG-GATES. Groups representing “Fatty Degeneration, Cellular infiltration, and Hydropic Degeneration”, and hypertrophy are the most frequently conserved. Group 1 represents cellular infiltrate (mixed cell and lymphocytic inflammatory cell) and glycogen accumulation. The amount of conservation on a module level is reflected by the joint histopathology meanings.

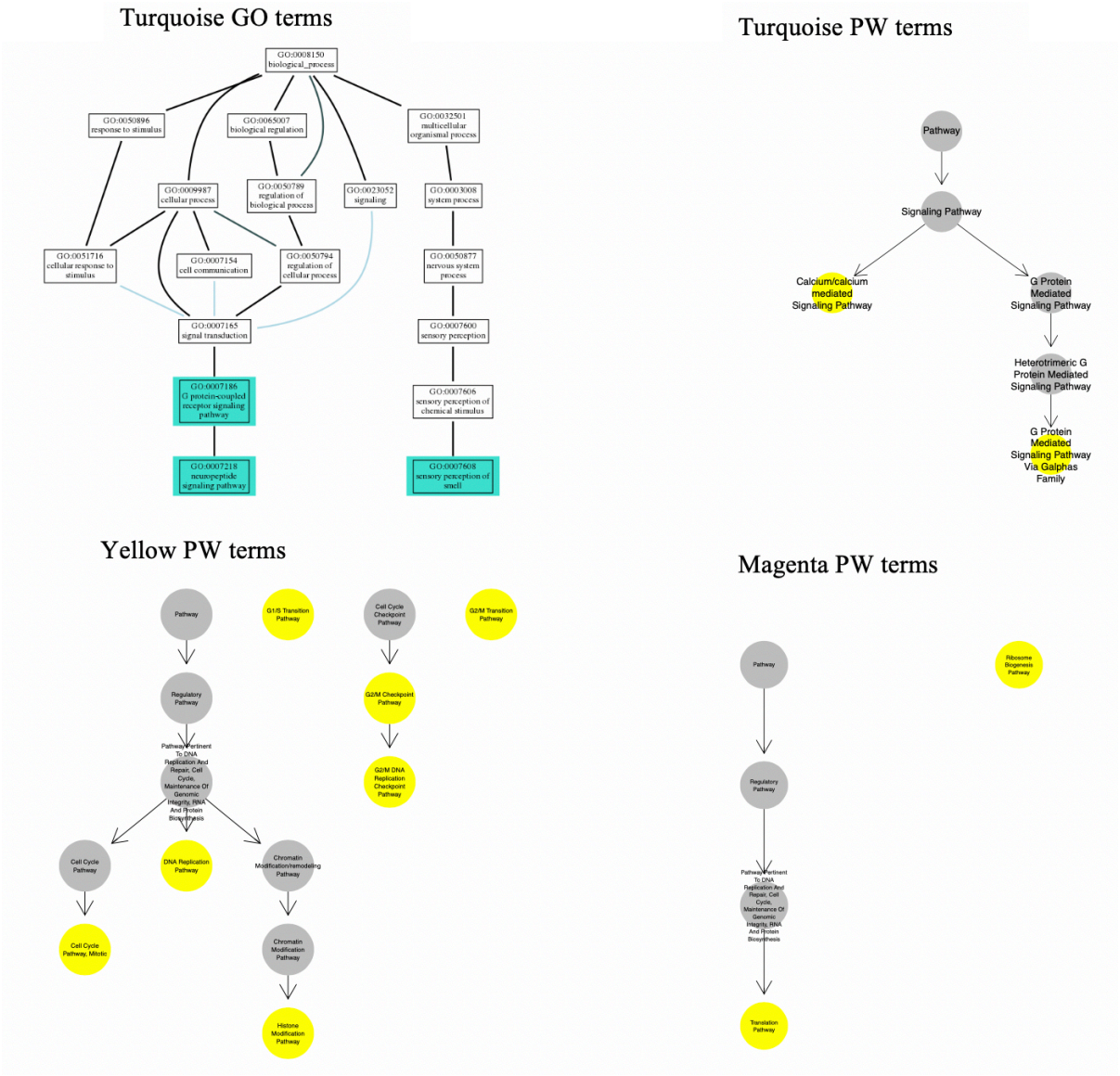



*Figure 3-9 Gene Ontology and Pathway ontology terms enriched in the yellow, magenta and turquoise modules (the brown module had no significantly enriched terms). It can be seen that these enriched terms show that a wide range of functional biological space is covered. PW means pathway.*

Open TG-GATEs contains histopathology observations that are not included in the definition of groups 1-6, and their biological meaning for "hypertrophy" and "hydropic degeneration" are briefly summarised. The former is the enlargement of an organ (liver in this case), that is caused by an increase in size of cells. When this is compound or drug-induced, fundamental cellular process have been shown to play a role.[147] These include "altered oxidative status, fatty acid metabolism, energy production and utilization, cell turnover and altered hepatocellular cytoplasmic, and nuclear morphology".[147–151] Of crucial importance, in this case, is the fatty acid metabolism. The method presented above clearly

shows its importance in Group 1. The latter, hydropic degeneration, is the ballooning in the size of a cell due an increase in the amount of water, as is similarly associated with other forms of hepatic degeneration. These histopathology observations are highly related between databases when considering gene expression space.

Group 1 is most conserved with histopathology signature definitions from Open TG-GATEs that show concordant meaning. This analysis was then performed over all the remaining signatures to determine the applicability of these across different groups.

#### 3.2.2.2 Summary of Groups 2-6

The same analysis on database concordance was performed for the remaining groups (groups 2-6). This concordance is expressed as the frequency of module conservation and is shown in Figure 3-8. It should be noted that Open TG-GATEs histopathology signature definitions are only represented here if they have at least one conserved module with one of the 6 toxic groups. The remaining groups represent a wider variation in histopathology and are not shown.

Group 2 (mixed infiltration and glycogen accumulation) is conserved with the signatures 'Fatty Degeneration, Cellular infiltration, and hydropic Degeneration' and 'hypertrophy'. There is no conservation with "glycogen deposit", as this is different from glycogen accumulation.[128,152,153]

Group 3 (mixed infiltration, glycogen accumulation and lymphocytic inflammatory cell infiltration) is defined by the same terms as Group 1, albeit a more severe finding. As such, it's conservation mirrors that of Group 1. The main difference, in terms of module conservation, between Groups 1 and 3 is the latter's conservation with 'cellular infiltration' and 'increased mitosis and eosinophilic granular degeneration' signatures. The former suggests that the increased amount of cellular infiltration observed in Group 3 (compared to Group 1), produces a signal that could be less dependent on the entire histopathology signature.

Group 4 (mixed infiltration, fatty change, glycogen accumulation, lymphocytic inflammatory cell infiltration) modules are most highly conserved with 'microgranuloma' and 'hypertrophy' signatures. Lipidic vacuolation (fatty change) frequently accompanies cell death but the mechanism of its role when induced by low molecular weight compounds is not known.[154] This method provides hypotheses on what might be occurring: Within Group 4's conserved modules, one has 'G protein-coupled receptor signalling pathway' (GO:0007186)

enrichment. It has been previously hypothesised that the local lipid environment plays a role in the signalling pathway, with some evidence to support this.[155] Further enquiry into this specific modules' hub genes may provide an insight into the lipidic vacuolation observation.

Group 5 (mixed infiltration, glycogen accumulation, lymphocytic cell infiltration and hepatocellular necrosis) uniquely has conserved modules with 'swelling'. Swelling is a key part of necrosis, and occurs after mitochondrial impairment, ATP deletion and the resultant failure of ion pumps.[156] It is not, however, conserved with the 'necrosis' group from Open TG-GATEs.

Group 6 (mixed infiltration, glycogen accumulation, lymphocytic cell infiltration and hepatocellular necrosis) is defined by the same terms as Group 5, with more severe terms in every observation. This, however, is not reflected entirely in their conservation with Open TG-GATEs groups. Only 'hypertrophy' (which is the most conserved term for Group 6 and second most for Group 5) and 'cellular infiltration' are conserved for both groups. Group 6 is additionally conserved with 'microgranuloma', 'increased mitosis and granular eosinophilic degeneration', 'increased mitosis' and 'necrosis' groups. The latter of these is within its definition, and therefore expected. Microgranuloma are a small collection of macrophages that occur typically with inflammatory cells (it is estimated that 13-33% have no discoverable aetiology in humans).[157–159]

The conservation of each Group's modules with TG histopathology signatures show direct and indirect associations. Direct associations are clear mappings of histopathology groups (e.g. cellular infiltration). Indirect associations, on the other hand, occur when the mappings are not as apparent, but these suggest the underlying relationship on a gene/molecular level as demonstrated with enrichments in ontology terms such as swelling and necrosis. Overall, we can conclude that unsupervised analysis of histopathology groups via co-expression networks enables their mapping to molecular processes that are DILI relevant.

### 3.2.3 PPI comparison

The interactions of genes within modules were compared to known interactions of the gene products. The size and significance of the resultant KS test can be seen in Table 3-4. From this, it can be concluded that the umber of interactions of gene products defined by modules is significantly more than that of random. This has accounted for the size effect of each module that was defined.

*Table 3-4 The results of the Kolmogorov-Smirnov test, used to determine if randomly selected proteins have the same distribution of interactions as the gene products of genes in all calculated modules. The test statistic and significant p value reveals they do not share the same distribution. This furthers a weight of evidence that there is biological value in the WGCNA approach to define modules.*

| Test statistic (D) | P value |
|---|---|
| 0.480 | $3.42 \times 10^{-5}$ |

## 3.3 Discussion

It is well known that gene expression data is often noisy, with batch effects from different laboratories, ambient conditions and a low number of biological and technical repeats. As such, external validation is crucial to assess the biological interpretation of individual microarrays. One common method to overcome this is normalisation with respect to housekeeping genes.[160] This is based on the assumption (and limitation) of there being genes who's expression is constant in a cell. This analyse is often performed with reverse transcription polymerase chain reaction experiments, but is rarely done in microarray technologies.[161] Research on applying this form of normalisation on microarray platforms found that it was most applicable where the expression/fold change are close to significance cut-offs.[161] As such, it is not suitable for the co-expression method used here.

This method of considering all the histopathological observations gives a holistic view of effects that compounds have. This is in direct contrast to previous work that chooses a specific histopathological assay of interest to determine its drivers.[98] Indeed, this view of using histopathology observations to guide and determine 'disease state' has recently shown to determine novel and known mechanisms of liver injury.[25] This histopathology-led approach, as taken in this work, treated *in vivo* toxicity as a systems biology problem. Here, this determined associations between databases that had not previously be found.

Following on from this systems approach, this investigation enabled the consistency evaluation of liver-specific histopathology signatures between DrugMatrix and Open TG-GATEs on the gene expression level. The signatures covered glycogen accumulation, cellular infiltration, necrosis and fatty change. The compounds within DrugMatrix were selected to

represent a wide area of therapeutic space. As such, they have fewer, less wide-ranging histopathology observations, compared to the toxicity-annotated compounds from Open TG-GATEs. Creating co-expression networks and determining gene-histopathology associations within DrugMatrix is more useful, as these can be validated in the wider-ranging histopathologies in Open TG-GATEs. The unique approach of using module conservation found large numbers of known associations to liver injury. Known pathways and GO terms have been found, however novel importance of particular genes (from hub genes) helps to widen hypotheses.

However, this method is not without its limitations. Firstly, it assumes that all of the histopathological observations are complete, with no other phenotype occurring. This is an inherent limitation of the data. Secondly, it suffers from the same limitations of a small sample size, which is exacerbated when considering samples' full histopathology profile. Here, the histopathology ontology (HPATH) was implemented to counter this hurdle, but it does require a balance to determine where on the ontology the grouping is made (as shown in Chapter 2). The balance is between the number of compound-dose-timepoint instances that allow for statistical significance to be reached and how specific the resultant histopathological profile is. The number of modules that are conserved between toxic groups that share one histopathology is small. This reflects the stated systems view that all histopathology observations must be considered together.

Pharmacological profiles from bioactivity clustering fail to discriminate toxicity. These profiles do not consider the absorption, distribution, metabolism and excretion (ADME) properties of the compounds, nor consider the exposures at the *in vivo* level. Additionally, they are limited in number and are not specific to rats. Thus, their failure is unsurprising. Determining differentially expressed genes and their pathway enrichments is one such method that is based directly upon the measured values. As such, and as shown above, these methods fail to discriminate between the toxic and non-toxic cases (Supplementary Figures 2 and 3).

Despite these limitations, there is much more biological information that is captured by framing the question more appropriately. WGCNA does not directly use the values measured in a microarray experiment, but instead, uses the correlation of these values between genes (or probes) to make use of the relationship between them. Based off the assumption that highly correlating genes are biologically related, its data driven approach acts to reduce the dimensionality of the dataset from ~14,000 genes to ~30 modules. The ability to compare and contrast different networks identifies the modules that are conserved or not

conserved between them in an unbiased manner. Interpreting these modules can be performed using gene, pathway, GO term, transcription factor and disease annotations, which confirm known associations and provide suggestions for further research.

The method goes on to suggest potential hypotheses for phenotypes: the involvement of lipid in G protein couple receptor pathways for efficient signalling, for example. It also suggests genes (from hub gene analysis) that have been annotated for specific functions, which concur with the defined histopathologies.

This work confirms previous, smaller studies: a case study on the toxic effect of carbon tetrachloride on liver.[162] Lipid binding and stress response at the gene expression level were associated with fatty degeneration and other histopathologies. This work widens the scope from specific compounds (with well characterised modes-of-action) to connecting signals between gene expression (from a wide range of compounds) to histopathologies. These associations help build up a set of "toxic modes-of-action", combining what is already known and suggesting previously unknown hypotheses.

## 3.4 Conclusion

Histopathological assays were used to create a systems view of toxicity. More traditional methods, determining differentially expressed genes and their enrichment, and pharmacological profile, failed to discriminate between the toxic and non toxic cases. Weighted gene co-expression network analysis revealed the transcriptomic level changes on a range of histopathology signatures. For example, glycogen accumulation, mixed infiltration and lymphocytic inflammatory cell infiltration is associated with fructose metabolism, gluconeogenesis and chemokine response pathways.

Conservation of modules between databases showed correlation of specific phenotypes, through the behaviour of groups of genes. Additional hypotheses were suggested for the presence of microgranuloma and lipidic vacuolation.

# 4. Concordance of transcriptomics data at different time points between similar and distinct rat liver histopathology observations based on the DrugMatrix and TG-GATEs databases

The determination of earlier gene expression signatures which pre-empt the phenotype will provide value in predictive models. This chapter aims to determine if a biologically meaningful signal is conserved across time points (1 day and 4/5 days). If late stage toxicity end points can be predicted by early time points, fewer compounds will fail due to toxicity later in the drug development pipeline. This precise question was asked using Open TG-GATEs and one day vs 28 day toxicity assays.[163] High levels of similarity were found in the gene expression profiles at both stages, based on the overlap of the top and bottom ranked genes. Specific histopathology observations and their occurrence and development were not the focus of any enquiry here.

The use of laser microdissection was used to determine the development of Helicobacter-induced mucosa-associated lymphoid tissue lymphoma in the gene expression domain.[164] In this study, the gene expression profiles of various histopathology stages of disease were determined. The time analysis determined genes coding for the immunoglobulins and the small proline-rich protein Sprr 2A to be critical for the initiation of reactive lymphocytes into the stomach. The subsequent step is characterized histologically by the antigen-driven proliferation and aggregation of B cells and the gradual appearance of lymphoepithelial lesions, in which the laminin receptor 1 and multidrug resistant channel MDR-1 are over expressed. This time-dependent analysis showed how the disease progressed on the gene expression level and this can equally be applied to toxicity. For example, the progression of disease (due to chronic dietary TCDD (a known toxicant) exposure) in zebra fish has been studied to determine the growth and development of histopathology at a gene expression level has been performed.[165] This work found dysregulated genes involved in pathways associated with cardiac necrosis/cell death, cardiac fibrosis, renal necrosis/cell death and liver necrosis/cell death , which are relevant to the endpoints discussed so far.

Time course, gene expression data are not easily modelled. A study that determined genes predictive of “DILI most concern” using Open TG-GATEs data found Time-series work has been performed modelling “did not perform well on directly measured gene expression values”.[166] To improve the situation, co-expression network methods have subsequently been

employed in this work to determine how networks change over time, in a novel manner. One such method used weighted gene co-expression network analysis across the entirety of DrugMatrix and Open TG-GATEs to determined gene – anchoring phenotype associations.[102] Time series analysis of 3 hour through to 14 hour gene expression data identified modules associated bile duct hyperplasia varying over time. Whilst this does allow for co-occurance of histopathology observations (necrosis, inflammation and fibrosis, in this case), it uses modules that were created in a network covering all time points and exposures. In this work, networks are built on separate histopathology signatures to allow the creation (or lack thereof) of specific modules in case. Separating out time, dose and histopathology signatures allows (as performed in this work) allows for both background biology and toxicity specific functions to be examined directly.

This work uses weighted gene co-expression network analysis (WGCNA) to determine the conservation between supervised clusters of histopathologies. Histopathology observations are not considered independent of each other and so signatures created across all observations. These map to toxic groups that consist of compound-dose-time point expression profiles. It goes on to introduce a metric, W, that can be used to address the similarities between networks, showing where and what the biological meaning of the similarity occurs (step 1 in Figure 4-1). Toxicity specific and generic conserved biological pathways were determined and examined in a data driven manor (step 2 in Figure 4-1). This was then validated against TG (step 3 in Figure 4-1), to test whether the network similarity value, W, connects matched histopathology observations on the gene expression level, between databases.

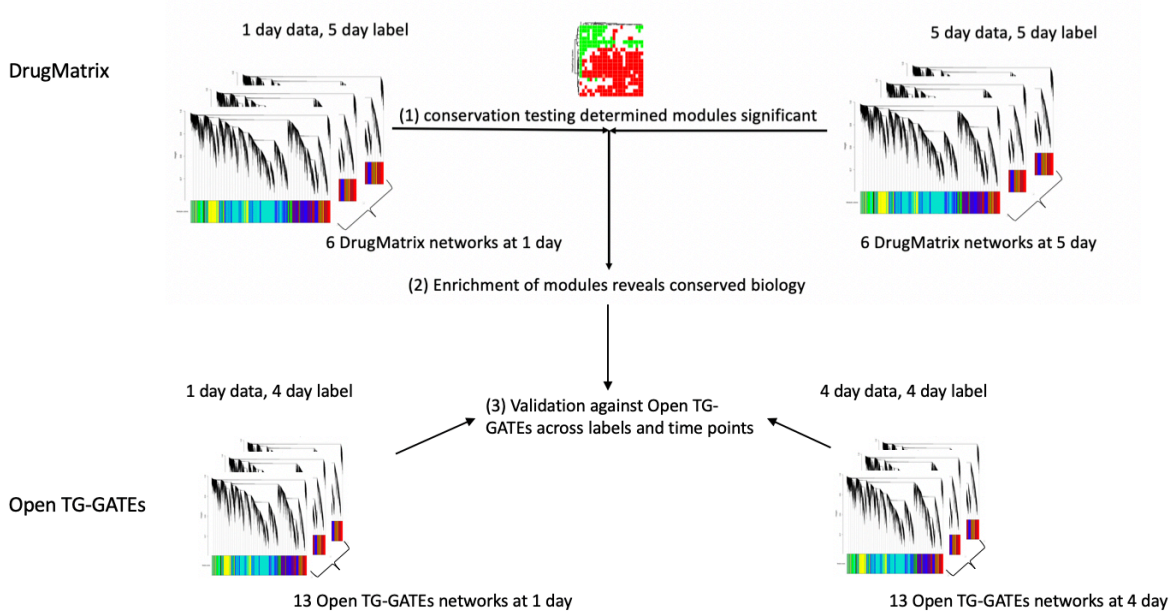



*Figure 4-1 The workflow for this study. Coexpression networks were created for six toxic compound groups (each of which caused identical histopathology phenotypes) based on the DrugMatrix data at both 1 and 5 days (12 networks in total). Their network modules were (1) tested to determine module conservations across time points. These were then (2) enriched with pathway and GO terms, allowing determination of time-independent biological conservation within the DrugMatrix data. The modules were (3) furthermore validated against two different sets of toxic groups from Open TG-GATEs (at 1 and 4 days).*

## 4.1 Methods

The histopathology signatures, toxic groups and gene expression profiles from the previous chapters were used in this one also, and their determination was identical. The 1 day time point data has not been until this point.

### 4.1.1 Weighted Gene Coexpression Network Analysis (WGCNA)

Weighted Gene Coexpression Network Analysis (WGCNA) was performed using the 'wgcna' package (version 1.66) in R.[80,110] Networks were created with softpower = 10, for each toxic group (Groups 1-6 at 5 day, Group1 1-6 at 1day, control and TG groups) on their raw, RMA-adjusted gene expression values. Modules were formed in each network using flashClust[167] and average linkage and a cut height = 0.99 for smaller modules. The remaining parameters were set as default.

Module conservation was determined by permutation testing of mean densities and correlations, both inter- and intra- modular, of connectivity, adjacency matrices and eigengene values.[83] 30 permutations were performed, with a maximum module size of 1,000 genes. Zsummary scores greater 10 were taken to show significant conservation, and less

than 2 shows no evidence of conservation.[83] A module consisting of randomly selected genes (n=1,000) was created 30 times and the same permutation test was applied to determine the false discovery rate of conservation scores. It has been shown both here and previously[83] that Zsummary score correlates linearly with modules size (Figure 4-2): the larger the module, the higher the chances of significant overlap. However, the median rank of the modules does not suffer from the same size dependency (Figure 4-2). The median rank is comparative between all of the modules within a network, and so is insufficient to solely determine whether a module is conserved or not. As such, a module can be considered significantly conserved if Zsummary >10 (as recommended[83]) and median rank < k * number of modules (to be determined). To determine a suitable value of k, values of k between 0 and 1 were selected at 0.1 intervals.

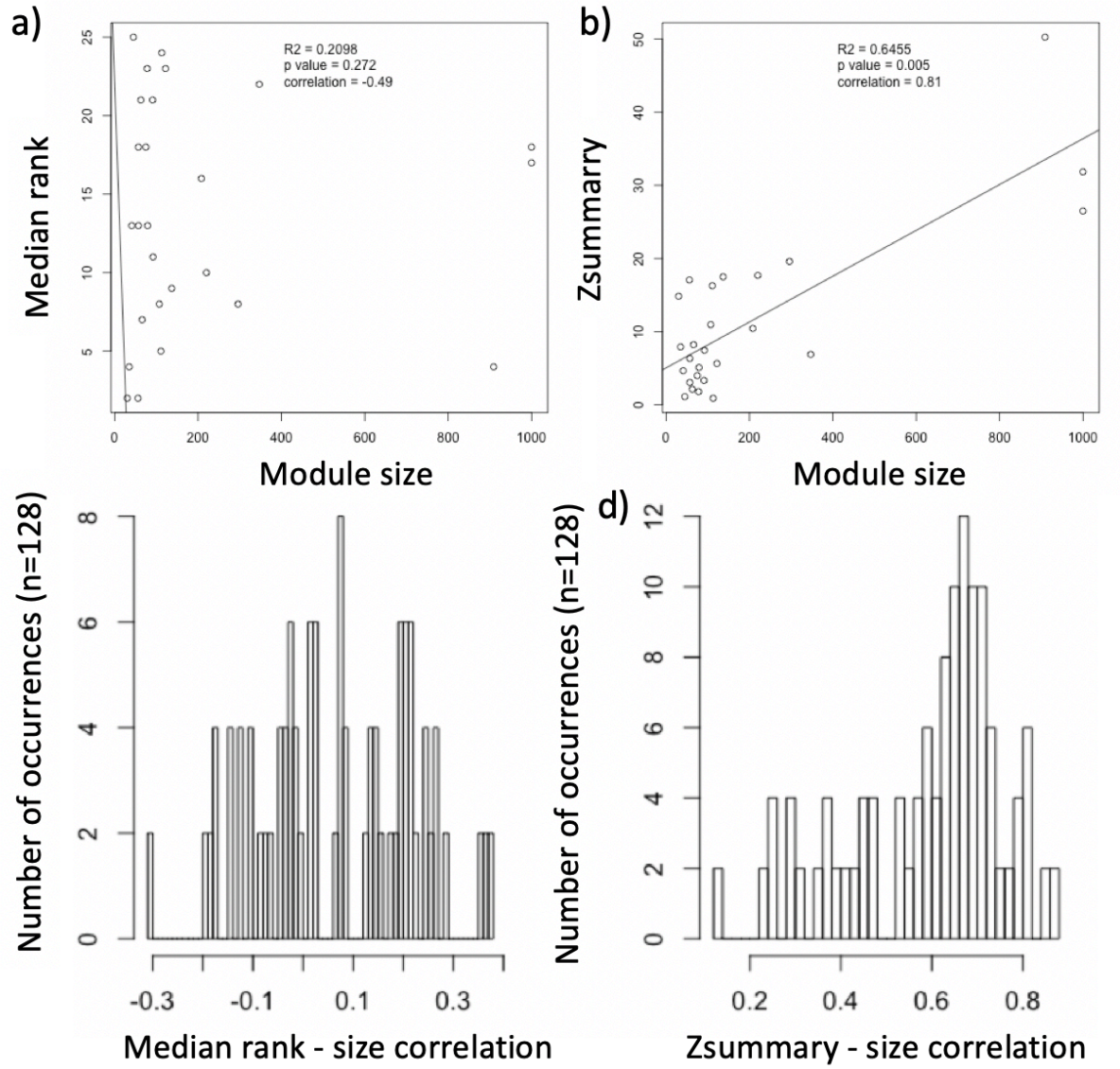



*Figure 4-2 Analysis of the effect that module size (number of genes) plays on the median rank (the average of the median module rank for 6 network characteristics [83]) and Zsummary score. (a) Correlation of median rank with modules size, with associated Pearson product moment correlation p value. A linear relationship cannot be seen. (b) Correlation of Zsummary with module size, which shows a linear relationship. (c) Shows the Pearson product moment correlation values for all tested modules for Group 1 between median rank and modules size. No strong relationship can be observed. (d) Correlation between Zsummary and module size, showing a positive correlation across the tested values. The Zsummary score, therefore, is not appropriate when determining if a module is conserved between networks. We conclude that joint cut-offs of median rank and Zsummary score can be used to determine is a module is conserved between networks, as it is in this work.*

In order to determine a suitable median rank cut off for networks with differing numbers of modules, the fraction of median rank is used. The total number of module conservations with a Zsummary score greater than 10 (the suggested significance cut off[83]) are shown in Figure 4-3. As shown, varying the median rank fraction changes the number of associations considered conserved. At high values (greater than 0.8), there is relatively little change, indicating that there are few modules with significant Zsummary scores and near the bottom of ranked conservations. In order to minimise false conservations, a value of 0.5 for a median rank cut off was selected.

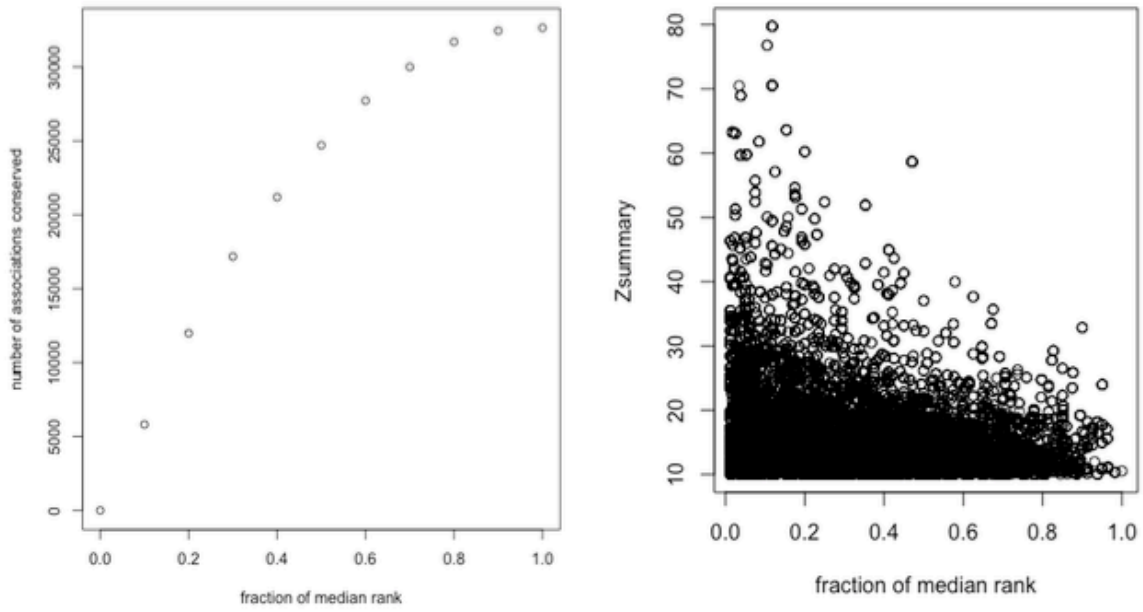



*Figure 4-3 (a) the relationship between a median rank cut-off and the number of associations (modules defined in one network being tested in another), and (b) between median rank cut-off and Zsummary score. All associations with high fraction value (i.e. are low in median rank list) represent false positives if only considering a Zsummary cut-off. Hence, in this analysis we chose to use a cut off of 0.5*

### 4.1.2 Network Conservation

The similarity and conservation of each network, W, was calculated using the fraction of the conserved number of modules against all modules:

$$W = \frac{Number\ of\ conserved\ modules}{Total\ number\ of\ modules} \quad (1)$$

As each network defined a set of modules, W is therefore directed. As such, the similarities of each group can be treated as the nodes and W as the directed edge weight. This was determined and viewed in Cytoscape.[168] K-medoids clustering was performed using ClusterMaker[169] with k = 6, using W as edge value.[169] Six clusters were formed, three of which contained "self transformations". Self transformations are defined as toxic groups from 1 day going to the same toxic group at 5 day (e.g. Group 1 1day -> Group 1 5 day, Table 4-1), whereas non-self transformations occur between different group labels.

*Table 4-1 definitions of self transformations within DrugMatrix.*

| **Transformation** | **self transformations** | | |
|---|---|---|---|
| T1 | Group 1 1day | -> | Group 1 5 day |
| T2 | Group 2 1day | -> | Group 2 5 day |
| T3 | Group 3 1day | -> | Group 3 5 day |
| T4 | Group 4 1day | -> | Group 4 5 day |
| T5 | Group 5 1day | -> | Group 5 5 day |
| T6 | Group 6 1day | -> | Group 6 5 day |

### 4.1.3 Module Enrichment and biological function overlap

Gene Ontology[62] terms and pathway annotations were downloaded from the Rat Genome Database.[58] Overrepresentation analysis was performed on each module using the Fisher exact test from the 'enrichr' package[170] against a background of all measured genes.[170] The FDR (Benjamini-Hochberg) multiple hypothesis correction[171] was applied on the resultant p-value and a corrected cutoff < 0.05 determined significance.

To determine the asymmetric biological overlap of two networks, each comparison was considered as a transformation (e.g. Group 1 1day data -> Group 1 5 day data). Each transformation had associated GO and pathway terms, defined as those enriched from the conserved modules. A Fisher's Exact Test was performed in R on the overlap of these terms when comparing two transformations (Table 4-2). The test results were visualised using heatmap.2 from the 'gplots' package in R.[172] Overlaps of enriched terms determined consensus toxic groups and time points, and unique terms were calculated using overlapper from the 'systemPipeR' package.[173]

*Table 4-2 Confusion matrix used to calculate the significance of overlap between two transformations on GO and pathway term levels. The Fisher's Exact Test, with FDR (Benjamini-Hochberg) corrected p value, was used to determine significance*

| **Confusion matrix** | Terms conserved in Transformation-1 | Terms not conserved in transformation-1 |
|---|---|---|
| Terms conserved in Transformation-2 | GO/pathway terms enriched in both transformations (i.e. enriched GO terms from conserved modules) | terms enriched in transformation-2 but not in transformation-1 |
| Terms not conserved in transformation-2 | terms enriched in transformation-1 but not in transformation-2 | Remaining GO/pathway terms |

## 4.2 Results and Discussion

### 4.2.1 Network comparison: time and histopathology signature dependence

We firstly analysed the concordance of co-expression networks in order to understand the relative influence of time and histopathology signature. This was performed by considering a "network of networks", to examine the networks relationships. Such a network consists of each toxic group as a node and the directed, weighted edge as the network conservation value, W (Figure 4-4). It can be seen that three of the six clusters (Groups 1, 3, and 4) form self transformation clusters. This implies that the histopathology signature which defines each toxic group has a greater effect on gene expression than the time at which the gene expression was measured.

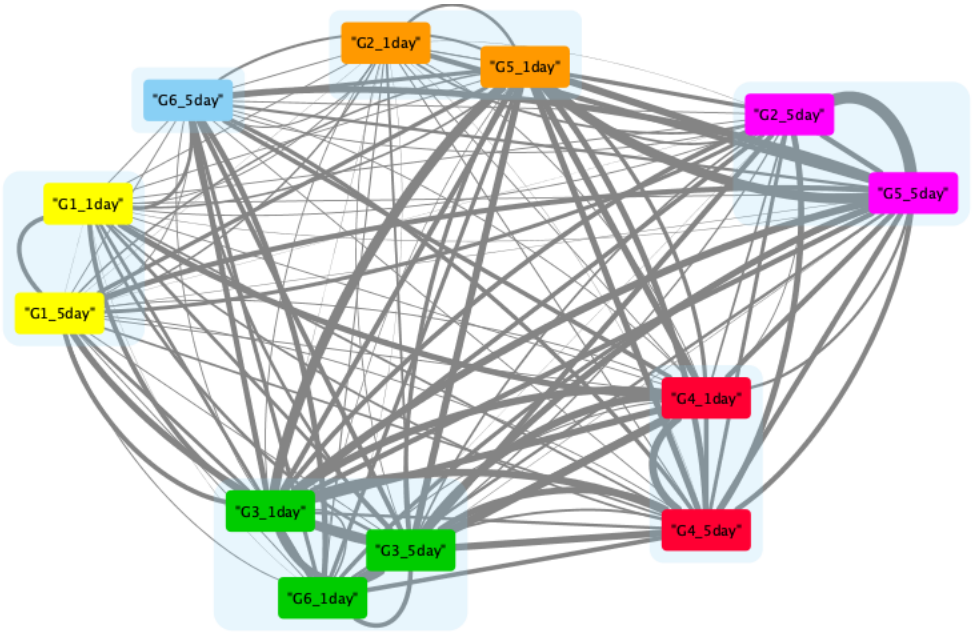



*Figure 4-4 Network of networks representation. Each node represents the co-expression network of a toxic group at a particular time point, and GX is Group X. Edges represent the directed weight, W. This shows that three out of the six clusters prioritise the toxicity label, over the time point. Two of the six prioritise the time point and one is (Group 6 at 5 day) is dissimilar to all.*

To investigate this further, we next analysed the importance of "self transformations" in each node, going from 1 day groups to 5 day groups. These distributions are shown in Figure 4-5. Four of the six of these weights are in the top 75$^{th}$ percentile. This observation,

and the clustering of the networks, was somewhat unexpected but implies the significance of toxic group membership when compared to time point of measurement, for the histopathology signatures studied here. As such, the biological overlap of this conservation was considered to determine the time independent facets of the underlying biology and toxic group specific terms.

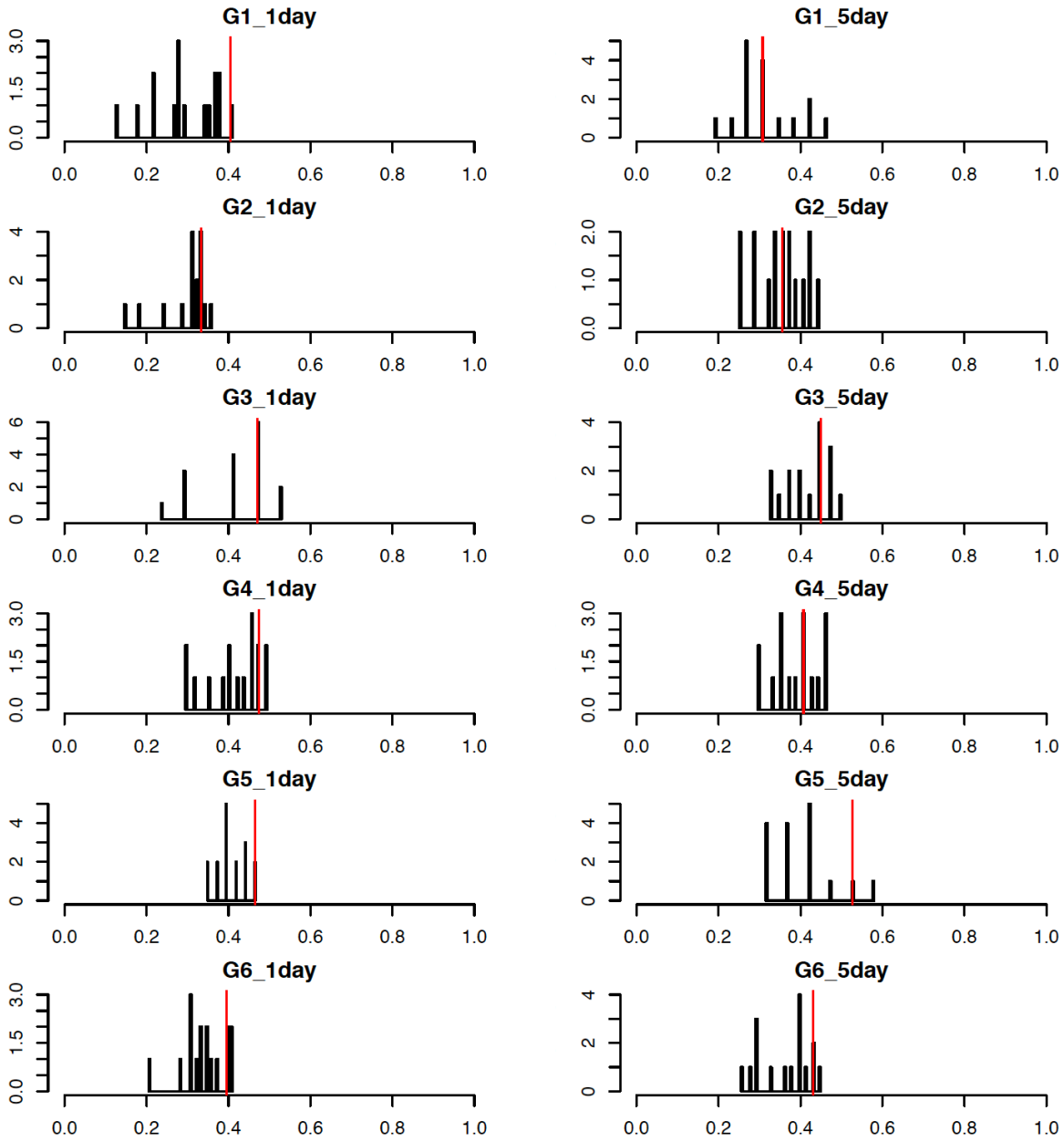



*Figure 4-5 the outgoing weight edge distribution for each group, at two time points. The red line is the weight edge for the same toxic group at a different time point. When considering the forward transformation (1 day to 5 day), this edge is in at least the 75th percentile of edges for 4 out of 6 cases. This means that there is significantly more conservation within a group of compounds with a shared histopathology definition across time points, than between compounds with distinct histopathology annotations.*

We next analysed the biological meaning of the similarity between networks in order to determine which pathways are conserved in each transformation (and so which pathways are time/label independent). Calculated using Fisher's Exact Test and the confusion matrix (Table 4-2), the results are shown in Figure 4-6. This shows the adjusted p value and that all transformations (except Group 2 1 day -> Group 2 5 five day (T2) GO terms) contain a statistically significant number of overlapping enriched terms. We conclude that the biological terms significantly conserved in each transformation are consistent across transformations. The precise values of these terms are shown in Supplementary Table 5

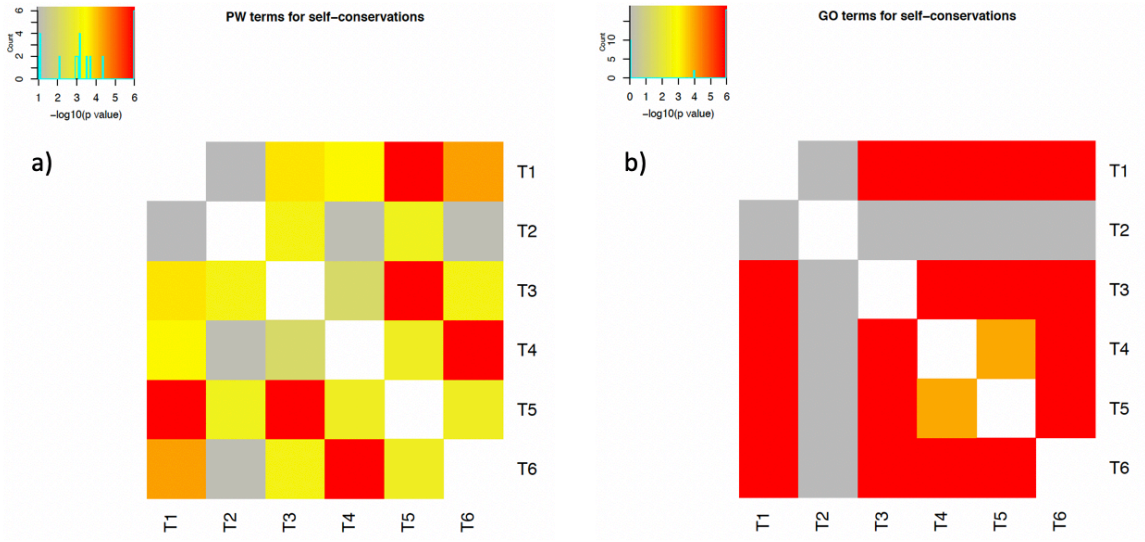



*Figure 4-6 the significance of overlapping pathway (PW; a) and GO terms (b) for the self transformations. The scale is represented as the -log(adjusted p value). T1 refers to Transformation 1 (i.e. Group 1 1day -> Group 1 5day). This shows that the pathway and GO terms that are significant in each transformation are statistically significant across the other transformations. Therefore, biological meaning in a transformation is significantly conserved across other transformations and so these are not label specific (the same terms are conserved in nearly each case). Only transformation 2 (Group 2 1 day -> Group 2 5 day) is not significant with the others.*

We next analysed the specific meaning of these terms, to determine their nature (background biology or toxicity specific). Significant pathway and GO term overlap showed some terms were conserved over the majority of transformations (Figure 4-7). "RNA transport pathway" was indeed conserved in each case. This is a regulatory pathway and only one of 93 associated diseases in the CTD refer to liver injury.[48] "translation" and "ribosome biogenesis" pathways are conserved in all except one transformation, and so these support the

hypothesis of underlying regulatory biology being most conserved. Similarly, for metabolic pathways, pyrimidine metabolism, is conserved in four of six. The remaining conserved pathway terms are related to background cell regulation biology, but terms with smaller overlap (e.g. in two transformation) include toxicity specific relationships. We conclude that the overlapped pathways conserved in each transformation represent more basic biological functions.


*Figure 4-7 Enriched pathways which overlap in at least 3 self-transformations. This shows the range of pathways that are conserved in the transformations; from S Aureus to general 'translation pathway' terms. This reflects the background biology which is not toxic group specific. The more specific terms are showed in the individual transformations (Supplementary Table 5) The colour represents which self transformations are included: blue is all transformations, green is Groups 1, 3, 4, 5, and 6, yellow is Groups 1, 4, 5 and 6, red is Groups 1, 3 and 6, and orange is Groups 2, 3, and 5. The pathway hierarchy comes from the Rat Genome Database..*

Groups 1 and 3 contain the same histopathology observations (mixed infiltration, lymphocytic inflammatory cell infiltration and glycogen accumulation), albeit with different severity scores. Their equivalent self-transformations (Group 1 1day -> Group 1 5 day, Group 3 1day -> Group 3 5 day) both contain the aminoacyl-tRNA biosynthetic pathway (Supplementary Table 5). This pathway consists of 38 genes, each coding for t-RNA synthetases. These are vital for the metabolic processes within cells. These are connected to various diseases, including cancers and metabolic diseases.[174] The connection between the

gene expression of Group 1, its defining histopathology and its relationship to cancers and metabolic diseases have been studied in detail.(Chapter 3) [131] Such a relationship could become vital in determining both the formation of histopathologies and their development to later stage toxicity read outs. We conclude the usefulness of this data led association, albeit with known relevance.

Similar relationships, however, do not exist within other groups; Groups 5 and 6 contain the same histopathology observations (hepatocellular necrosis, mixed infiltration, lymphocytic inflammatory cell infiltration and glycogen accumulation) but do not hold unique conserved pathways in their self-transformation. This is not easily explainable, however it is also not a finding in itself. From this we conclude that not all expected observations are shown in the gene expression data.

GO terms reflect the same biological conservations with "translation" and "ribosomal large subunit biogenesis" conserved across all transformations (Supplementary Table 5). Additionally, "biological process" is conserved in half of the self-transformations. This generic term confirms that the conserved signal represents basic biology. Groups 1 and 3 share the "liver regeneration" term, which is associated with the toxicity being considered in this study. These are fairly generic observations and do not add any biological meaning. From this, we can conclude that the method generates related pathways, but those that overlap in all transformations are non-specific.

### 4.2.2 Network Validation: internal DrugMatrix

We next analysed the internal validation with DrugMatrix to determine the significance of the pathway terms conserved in self transformations. Their significance must be determined otherwise their inclusion be an artefact of the data. The significance of these transformations were firstly determined against all non-self-transformations.

Significance within DrugMatrix between self transformations and all other transformations was determined by comparing enriched GO and pathways terms that were conserved between conserved modules from different toxic group networks. This is visualised (Figure 4-8) and the relative proportion of significant terms compared (Table 4-3). For pathways, there is a significant difference in the proportion of conserved terms between self-transformations, ranging between 16-45%. There are, therefore, a greater number of conserved pathways in the self transformations, adding support to the evidence that the toxic group definition plays a greater role in determining similarity than the time effects. From this,

we conclude the greater significance of histopathology signature (compared to time) in co-expression networks.

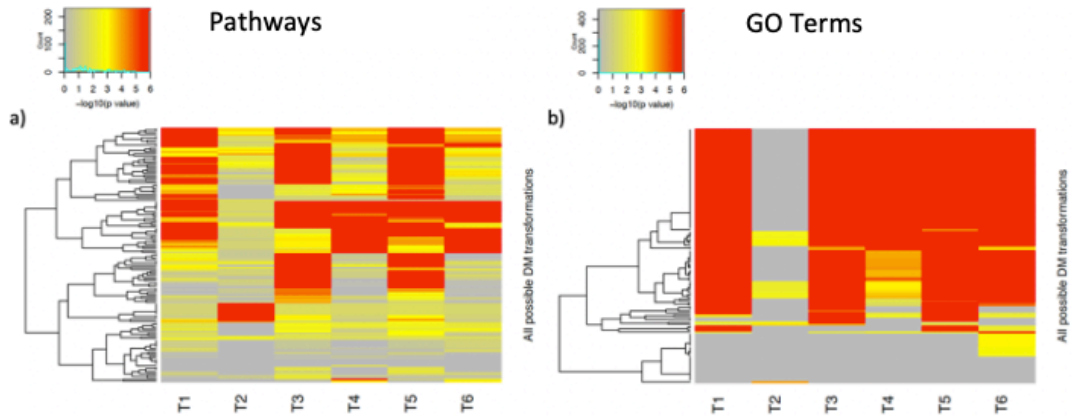



*Figure 4-8 significance of overlapping pathways (a) and GO terms (b) in pairwise comparison of self-transformations vs all possible transformation. This shows the relative importance of significant overlap between self-transformations and all possible transformations. There is a difference with enriched pathways, but this is not seen with GO terms. T1 is transformation one (Group 1 1 day -> Group 1 5 day).*

On the other hand, there is no significant difference in GO term conservations between transformations. This may reflect the lack of specificity and proportionately high number of GO terms (6,000 GO terms compared to 2,635 pathway terms). This redundancy in GO terms has previously been studied.[175,176] This analysis does not clarify the utility of using co-expression networks.

*Table 4-3 the proportion of terms conserved in comparing pathway (a) and GO (b) terms. The relative number show a large difference between self and all transformations for Pathway terms. There is no difference between the proportion of enriched GO terms between the self and all transformations. This analysis shows that the self-transformations have a disproportionate percentage of enriched and conserved pathways, compared to all other transformations. However, the same is not see with GO terms*

| A | Pathway | | | | | | |
|---|---|---|---|---|---|---|---|
| | **# significant of overlap between self-transformations** | **total # self transformation** | **Proportion (significant self transformations / total number)** | **# significant of overlap between all transformations** | **total # all transformations** | **Proportion (significant in all transformations / total number)** | **difference between proportions** |
| **G1 1 day->G1 5 day** | 4 | 5 | 0.8 | 73 | 132 | 0.55 | **0.25** |
| **G2 1 day->G2 5 day** | 2 | 5 | 0.4 | 22 | 132 | 0.17 | **0.23** |
| **G3 1 day->G3 5 day** | 4 | 5 | 0.8 | 84 | 132 | 0.64 | **0.16** |
| **G4 1 day->G4 5 day** | 3 | 5 | 0.6 | 48 | 132 | 0.36 | **0.24** |
| **G5 1 day->G5 5 day** | 5 | 5 | 1 | 97 | 132 | 0.73 | **0.27** |
| **G6 1 day->G6 5 day** | 4 | 5 | 0.8 | 46 | 132 | 0.35 | **0.45** |

| B | GO term | | | | | | |
|---|---|---|---|---|---|---|---|
| | **# significant of overlap between self-transformations** | **total # self transformation** | **Proportion (significant self transformations / total number)** | **# significant of overlap between all transformations** | **total # all transformations** | **Proportion (significant in all transformations / total number)** | **difference between proportions** |
| **G1_1day->G1 5 day** | 4 | 5 | 0.8 | 101 | 132 | 0.77 | **0.03** |
| **G2_1day->G2 5 day** | 0 | 5 | 0 | 11 | 132 | 0.08 | **-0.08** |
| **G3_1day->G3 5 day** | 4 | 5 | 0.8 | 102 | 132 | 0.77 | **0.03** |
| **G4_1day->G4 5 day** | 4 | 5 | 0.8 | 90 | 132 | 0.68 | **0.12** |
| **G5_1day->G5 5 day** | 4 | 5 | 0.8 | 103 | 132 | 0.78 | **0.02** |
| **G6_1day->G6 5 day** | 4 | 5 | 0.8 | 105 | 132 | 0.80 | **0.00** |

### 4.2.3 Network Validation: Open TG GATEs

We next investigated how well the co-expression network comparison works on external data, using the Open TG-GATEs database. For any gene expression-based work, the practical value depends on whether the signal found can be replicated across datasets and databases.[177] This external validation investigated three main points: total amount of conservation, group similarity, and comparison of similar histopathologies between databases.

Firstly, 1 day networks had a higher amount of conservation with the 1 day networks from TG compared to the 4 day networks from Open TG-GATEs (

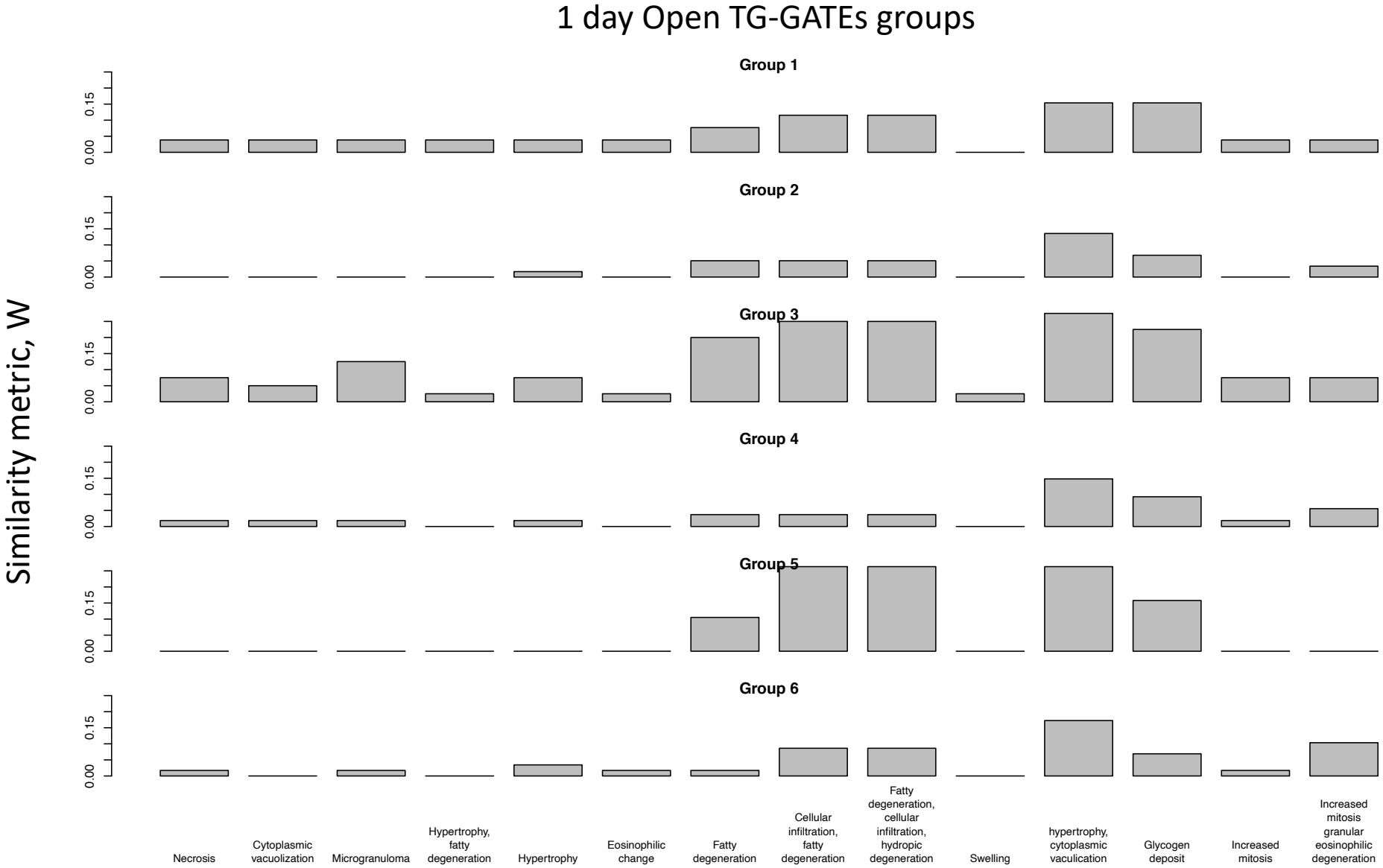
Figure *4-9*). This implies that, when comparing databases, gene expression data at 1 day is more similar to other 1 day data, at the network level. This is in line with expectations, as TG does not contain identical groups to those determined from DrugMatrix.

Secondly, there are significant similarities between Groups 5 and 6 at the 1 day stage (

Figure *4-9*), with conservation with identical Open TG-GATEs histopathology groups. Both of these are defined by 'mixed infiltration', 'glycogen accumulation', 'necrosis' and 'lymphocytic inflammatory cell infiltration'. Likewise, they are both solely conserved with the same 6 groups from Open TG-GATEs. These concern 'mixed infiltration', 'glycogen deposition', 'fatty degeneration' and 'hypertrophy'. Clearly, these are related to Groups 5 and 6 and so their conservation is expected. However, the TG group 'necrosis' was not conserved, despite it being present in Groups 5 and 6. This is unexpected but may be due to the Open TG-GATEs 'necrosis' observation occurring in insolation, compared to the signature of terms in the other groups. This may support the use of histopathology signatures.

Thirdly, of all the possible histopathology groups in TG, only those similar observations to Groups 1-6 have any conservation (

Figure *4-9* and Supplementary Table 1). This is encouraging, concerning the concordance of gene expression data across databases in the efforts towards creating predictive models of later stage toxicity.


1 day Open TG-GATEs groups
Similarity metric, W
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
0.00
0.15
Necrosis
Cytoplasmic vacuolization
Microgranuloma
Hypertrophy, fatty degeneration
Hypertrophy
Eosinophilic change
Fatty degeneration
Cellular infiltration, fatty degeneration
Fatty degeneration, cellular infiltration, hydropic degeneration
Swelling
hypertrophy, cytoplasmic vaculication
Glycogen deposit
Increased mitosis
Increased mitosis granular eosinophilic degeneration

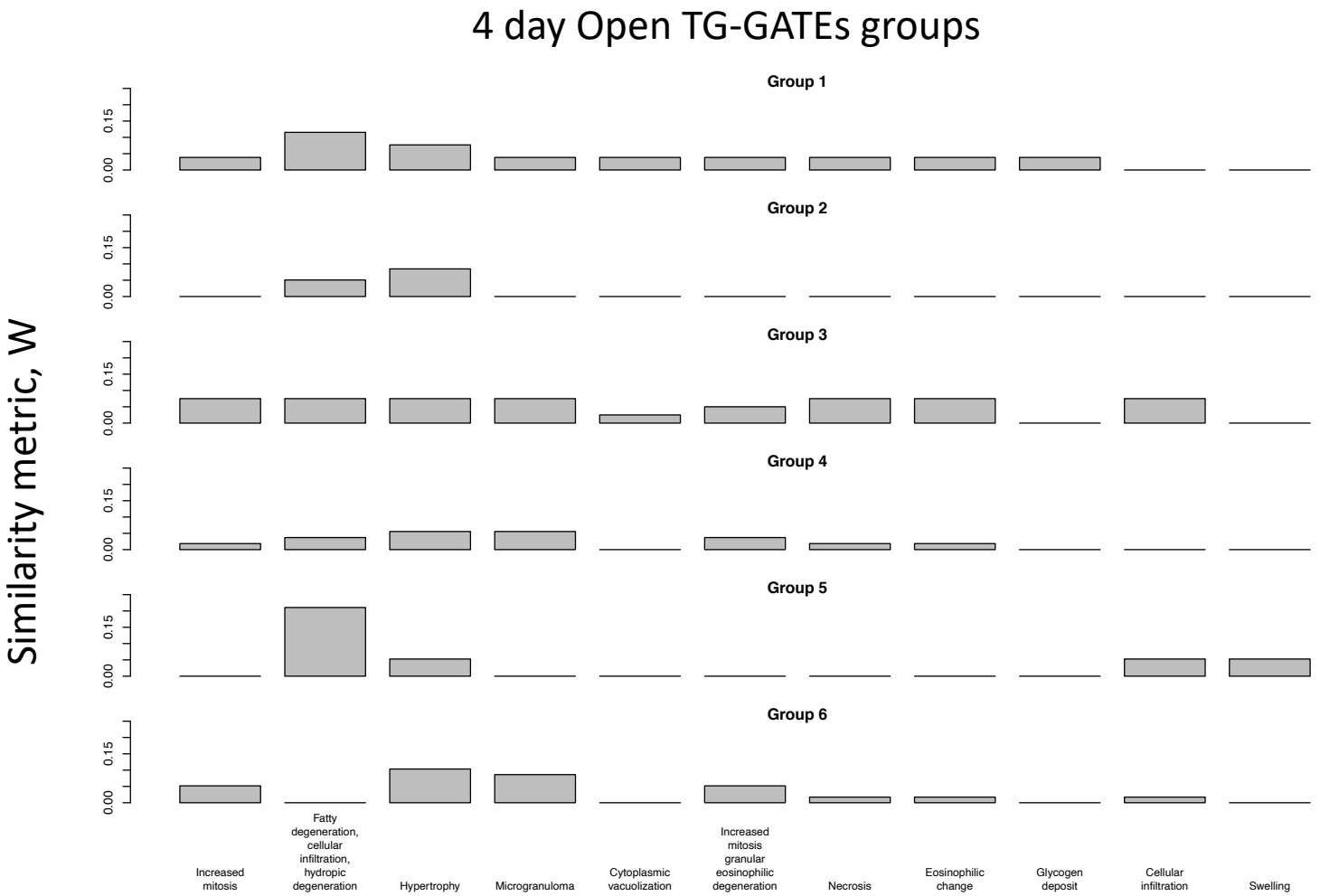



*Figure 4-9 The network similarity value, W, for the DrugMatrix toxic groups (Groups 1-6 at 1 day) compared to Open TG-GATEs groups at 1 and 4 days. This crucially shows higher W for the 1 day groups (the same time point as the comparison) and the similarity on a histopathology level, determined from a co-expression network level. Groups 1 and 3 and Groups 5 and 6 show very similar behaviour at 1 day, reflecting their similar histopathology signature. This figure shows the co-expression validation across databases, and that co-expression network similarity captures gene expression signal for phenotype conservation.*

## 4.3 Conclusions

This work determined differences in the effect that time and toxicity group definition have on gene expression data. Using DM, a systems level view of histopathology observations was used to determine histopathology signature, creating toxic groups. Networks built on these were compared to each other, using a novel metric based on conserved modules. These revealed that toxicity group membership has a more significant role in affecting gene expression than time point of measurement within a database. This encourages the development of predictive models from earlier data, modelling later stage data, in particular, it confirms the Sutherland *et al.* work that predicted later stage histopathology observations from earlier time point gene expression.[102] This analysis also leads to a greater understanding about what biology is conserved, and how specific it is to toxicity and underlying biology for compound-induced liver injury – specially glycogen accumulation, cellular infiltration, hepatocellular necrosis, and fatty change. Underlying terms include RNA transport, ribosome biogenesis, pyrimidine metabolism and translation. Toxicity specific terms suggest the role that aminoacyl tRNA synthesis has in metabolic processes, leading to glycogen accumulation and cellular infiltration. These network methods were confirmed using Open TG-GATEs, whose histopathology groups matched those from DM. Between databases, higher levels of concordance were found for networks built on gene expression data from the same time.

This work lays the groundwork for predictive modelling, by including biological knowledge-based information for supervised and semi-supervised methods. Additionally, it provides data-driven information about the precise role of tRNA in hepatoxicity and provides a metric for co-expression network comparisons to determine the effect of toxic group and time on conservation.

# 5. Conclusions

In conclusion, data-driven analysis of DrugMatrix and Open TG-GATEs led to the creation of novel histopathology signatures which define toxic groups (Chapter 2). The gene expression data measured concurrently to the histopathology phenotype revealed known and novel gene-phenotype associations (Chapter 3). The significance of these labels was determined in time-series analysis where the label was shown to be more significant than the time point of measurement on affecting gene co-expression networks (Chapter 4). These conclusions are valid for the five main histopathology observations: glycogen accumulation, mixed infiltration, lymphocytic inflammatory cell infiltration, hepatocellular necrosis and fatty change.

This work determined the concordance between these forms of data. It is, however, not without limitations. These include the noise within microarray platforms, dose dependency, and species specificity. The co-expression network methods used here (WGCNA) is weighted and so the relative impact of the noise is reduced. Biological noise (such as a stress reaction within the cell) was observed and cannot be discounted.

A key draw back here, and more generally in gene expression data, is the requirement of *a priori* knowledge in order to understand the gene expression. Known biological pathways, gene – disease relationships, and gene – transcription factor associations were required to determine the meaning of the data driven modules that were formed here. Key genes (e.g. hub genes) were suggested by the methodology and do provide a hypothesis for further experimentation. The results presented here do not offer a direct causation or proven biological pathway as output.

Dose dependency is the fundamental principle in toxicology and gene expression profiles of the same compound at different dose levels highly variable. As such, care must be taken when determining gene-phenotype associations and determining conservation between differing toxic groups. As was the case here where imatinib and bithionol was both present to two different groups each: imatinib was in group 2 (at 15 mg/kg) and group 5 (at 150 mg/kg), bithionol was in group 5 (at 59 mg/kg) and group 6 (at 333 mg/kg). It was shown that groups 5 and 6 were similar at the histopathological and gene expression level. However, group 2 was distinct from group 5 (and 6). The role of dose is clearly important. Future work with co-expression networks could determine which modules were dose dependent and to what degree, in a manner similar to the timepoint dependency in Chapter 4. However, this does

depend on the amount of data available, as the co-expression networks are more robust and reliable the higher the number of gene expression profiles that the networks are built upon.

Model species are also an obstacle in the field of toxicogenomics. This work has been based on rats but the manner in which a rat liver responds to compound exposure is not necessarily the same as the way that a human liver would respond. The same is true when considering the *in vivo/in vitro* comparison. However, an obvious extension to this work is to determine the network similarities and differences between gene expression from rats to rat and human *in vitro* hepatocytes.

# 6. Bibliography


1. S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson, *Health Policy*, 2011, **100**, 4–17.
2. J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, *Br. J. Pharmacol.*, 2011, **162**, 1239–1249.
3. J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, *J. Health Econ.*, 2016, **47**, 20–33.
4. K. Earm and Y. E. Earm, *Integrative Medicine Research*, 2014, **3**, 211–216.
5. J. Arrowsmith and P. Miller, *Nat. Rev. Drug Discov.*, 2013, **12**, 569.
6. D. Demortain, *Health. Risk Soc.*, 2008, **10**, 37–51.
7. N. Vargesson, in *Reproductive and developmental toxicology*, Elsevier, 2011, pp. 395–403.
8. I. J. Onakpoya, C. J. Heneghan, and J. K. Aronson, *BMC Med.*, 2016, **14**, 10.
9. R. K. Harrison, *Nat. Rev. Drug Discov.*, 2016, **15**, 817–818.
10. M. J. Waring, J. Arrowsmith, A. R. Leach, P. D. Leeson, S. Mandrell, R. M. Owen, G. Pairaudeau, W. D. Pennie, S. D. Pickett, J. Wang, O. Wallace, and A. Weir, *Nat. Rev. Drug Discov.*, 2015, **14**, 475–486.
11. S. U. Shah, *IOSR J. Pharm. Biol. Sci.*, 2012, **1**, 43–54.
12. S. Y. Kim and A. Moon, *Biomol. Ther. (Seoul)*, 2012, **20**, 268–272.
13. F. Svensson, A. Zoufir, S. Mahmoud, A. M. Afzal, I. Smit, K. A. Giblin, P. J. Clements, J. T. Mettetal, A. Pointon, J. S. Harvey, N. Greene, R. V. Williams, and A. Bender, *Chem. Res. Toxicol.*, 2018, **31**, 1119–1127.
14. H. Jaeschke, G. J. Gores, A. I. Cederbaum, J. A. Hinson, D. Pessayre, and J. J. Lemasters, *Toxicol. Sci.*, 2002, **65**, 166–176.
15. K. T. Suk and D. J. Kim, *Clin Mol Hepatol*, 2012, **18**, 249–257.
16. K. Köck, B. C. Ferslew, I. Netterberg, K. Yang, T. J. Urban, P. W. Swaan, P. W. Stewart, and K. L. R. Brouwer, *Drug Metab. Dispos.*, 2014, **42**, 665–674.
17. S. N. Ahmed and Z. A. Siddiqi, *Seizure*, 2006, **15**, 156–164.
18. S. Russmann, G. A. Kullak-Ublick, and I. Grattagliano, *Curr. Med. Chem.*, 2009, **16**, 3041–3053.
19. M. Chen, A. Suzuki, J. Borlak, R. J. Andrade, and M. I. Lucena, *J. Hepatol.*, 2015, **63**, 503–514.
20. H. Jaeschke, M. R. McGill, and A. Ramachandran, *Drug Metab Rev*, 2012, **44**, 88–106.
21. J. J. Xu, P. V. Henstock, M. C. Dunn, A. R. Smith, J. R. Chabot, and D. de Graaf, *Toxicol. Sci.*, 2008, **105**, 97–105.
22. D. Xu, M. Xu, S. Jeong, Y. Qian, H. Wu, Q. Xia, and X. Kong, *Front. Pharmacol.*, 2018, **9**, 1428.
23. D. Han, L. Dara, S. Win, T. A. Than, L. Yuan, S. Q. Abbasi, Z.-X. Liu, and N. Kaplowitz, *Trends Pharmacol. Sci.*, 2013, **34**, 243–253.
24. C. Pauli-Magnus and P. J. Meier, *Hepatology*, 2006, **44**, 778–787.
25. K. Shimada and T. J. Mitchison, *Mol. Syst. Biol.*, 2019, **15**, e8636.
26. H. A. Alturkistani, F. M. Tashkandi, and Z. M. Mohammedsaleh, *Glob. J. Health Sci.*, 2015, **8**, 72–79.
27. M. Cases, K. Briggs, T. Steger-Hartmann, F. Pognan, P. Marc, T. Kleinöder, C. H. Schwab, M. Pastor, J. Wichard, and F. Sanz, *Int. J. Mol. Sci.*, 2014, **15**, 21136–21154.
28. R. R. Maronpot, *Liver - Inflammation In: Cesta MF, Herbert RA, Brix A, Malarkey DE, Sills RC (Eds.), National Toxicology Program Nonneoplastic Lesion Atlas.*, 2014.
29. J. J. Hornberg, M. Laursen, N. Brenden, M. Persson, A. V. Thougaard, D. B. Toft, and T. Mow, *Drug Discov. Today*, 2014, **19**, 1131–1136.
30. S. Loiodice, A. Nogueira da Costa, and F. Atienzar, *Drug Chem Toxicol*, 2019, **42**, 113–121.
31. I. Manousaridis, S. Mavridou, S. Goerdt, M. Leverkus, and J. Utikal, *J. Eur. Acad. Dermatol.*

*Venereol.*, 2013, **27**, 11–18.
32. R. Gutzmer, J. C. Becker, A. Enk, C. Garbe, A. Hauschild, M. Leverkus, G. Reimer, R. Treudler, A. Tsianakas, C. Ulrich, A. Wollenberg, and B. Homey, *J Dtsch Dermatol Ges*, 2011, **9**, 195–203.
33. N. A. Meanwell, *Chem. Res. Toxicol.*, 2011, **24**, 1420–1456.
34. National Research Council (US) Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology, *Applications of toxicogenomic technologies to predictive toxicology and risk assessment*, National Academies Press (US), Washington (DC), 2007.
35. P. Joseph, *Food Chem. Toxicol.*, 2017.
36. J. A. Bourdon-Lacombe, I. D. Moffat, M. Deveau, M. Husain, S. Auerbach, D. Krewski, R. S. Thomas, P. R. Bushel, A. Williams, and C. L. Yauk, *Regul Toxicol Pharmacol*, 2015, **72**, 292–309.
37. R. Bumgarner, *Curr. Protoc. Mol. Biol.*, 2013, **Chapter 22**, Unit 22.1.
38. R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy, *J. Pharm. Bioallied Sci.*, 2012, **4**, S310-2.
39. S. Draghici, P. Khatri, A. C. Eklund, and Z. Szallasi, *Trends Genet.*, 2006, **22**, 101–109.
40. MAQC Consortium, L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker, *Nat. Biotechnol.*, 2006, **24**, 1151–1161.
41. M. S. Rao, T. R. Van Vleet, R. Ciurlionis, W. R. Buck, S. W. Mittelstadt, E. A. G. Blomme, and M. J. Liguori, *Front. Genet.*, 2018, **9**, 636.
42. W. R. Buck, J. F. Waring, and E. A. Blomme, *Methods Mol. Biol.*, 2008, **460**, 23–44.
43. W. M. Haschek, C. G. Rousseaux, and M. A. Wallig, *Haschek and rousseaux's handbook of toxicologic pathology*, Elsevier, 3rd edn., 2013.
44. B. Ganter, R. D. Snyder, D. N. Halbert, and M. D. Lee, *Pharmacogenomics*, 2006, **7**, 1025–1044.
45. Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani, and H. Yamada, *Nucleic Acids Res.*, 2015, **43**, D921-7.
46. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, *Science*, 2006, **313**, 1929–1935.
47. A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F.

Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F. Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W.-N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, and T. R. Golub, *Cell*, 2017, **171**, 1437–1452.e17.

48. A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, *Nucleic Acids Res.*, 2017, **45**, D972–D978.
49. A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, *Nucleic Acids Res.*, 2019, **47**, D948–D954.
50. R. Edgar, M. Domrachev, and A. E. Lash, *Nucleic Acids Res.*, 2002, **30**, 207–210.
51. N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma, *Nucleic Acids Res.*, 2015, **43**, D1113-6.
52. Q. Duan, C. Flynn, M. Niepel, M. Hafner, J. L. Muhlich, N. F. Fernandez, A. D. Rouillard, C. M. Tan, E. Y. Chen, T. R. Golub, P. K. Sorger, A. Subramanian, and A. Ma'ayan, *Nucleic Acids Res.*, 2014, **42**, W449-60.
53. L. Cheng and L. Li, *CPT Pharmacometrics Syst. Pharmacol.*, 2016, **5**, 588–598.
54. A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N. A. Fonseca, R. Petryszak, I. Papatheodorou, U. Sarkans, and A. Brazma, *Nucleic Acids Res.*, 2019, **47**, D711–D715.
55. A. Musa, S. Tripathi, M. Dehmer, and F. Emmert-Streib, *Front. Genet.*, 2019, **10**, 557.
56. E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander, *Nucleic Acids Res.*, 2011, **39**, D685-90.
57. D. Türei, T. Korcsmáros, and J. Saez-Rodriguez, *Nat. Methods*, 2016, **13**, 966–967.
58. M. Shimoyama, J. De Pons, G. T. Hayman, S. J. F. Laulederkind, W. Liu, R. Nigam, V. Petri, J. R. Smith, M. Tutaj, S.-J. Wang, E. Worthey, M. Dwinell, and H. Jacob, *Nucleic Acids Res.*, 2015, **43**, D743-50.
59. B. Alexander-Dann, L. L. Pruteanu, E. Oerton, N. Sharma, I. Berindan-Neagoe, D. Módos, and A. Bender, *Mol. Omics*, 2018, **14**, 218–236.
60. M. Kutmon, A. Riutta, N. Nunes, K. Hanspers, E. L. Willighagen, A. Bohler, J. Mélius, A. Waagmeester, S. R. Sinha, R. Miller, S. L. Coort, E. Cirillo, B. Smeets, C. T. Evelo, and A. R. Pico, *Nucleic Acids Res.*, 2016, **44**, D488-94.
61. R. A. Haw, D. Croft, C. K. Yung, N. Ndegwa, P. D'Eustachio, H. Hermjakob, and L. D. Stein, *Database (Oxford)*, 2011, **2011**, bar031.
62. Gene Ontology Consortium, *Nucleic Acids Res.*, 2015, **43**, D1049-56.
63. R. P. Huntley, T. Sawford, M. J. Martin, and C. O'Donovan, *Gigascience*, 2014, **3**, 4.
64. M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, *Nucleic Acids Res.*, 2014, **42**, D199-205.
65. A. Krämer, J. Green, J. Pollard, and S. Tugendreich, *Bioinformatics*, 2014, **30**, 523–530.
66. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, *Cell Syst.*, 2015, **1**, 417–425.
67. A. Anjum, S. Jaggi, E. Varghese, S. Lall, A. Bhowmik, and A. Rai, *J. Comput. Biol.*, 2016, **23**, 239–247.
68. D. Colquhoun, *R. Soc. Open Sci.*, 2014, **1**, 140216.
69. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, *Nucleic Acids Res.*, 2015, **43**, e47.
70. W. Shi, A. Bugrim, Y. Nikolsky, T. Nikolskya, and R. J. Brennan, *Toxicol. Mech. Methods*,

2008, **18**, 267–276.
71. F. Sirci, F. Napolitano, S. Pisonero-Vaquero, D. Carrella, D. L. Medina, and D. di Bernardo, *npj Syst. Biol. Appl.*, 2017, **3**, 23.
72. J. J. Babcock, F. Du, K. Xu, S. J. Wheelan, and M. Li, *PLoS One*, 2013, **8**, e69513.
73. P. Khatri, M. Sirota, and A. J. Butte, *PLoS Comput. Biol.*, 2012, **8**, e1002375.
74. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 15545–15550.
75. K. H. Jung, J. K. Kim, J. H. Noh, J. W. Eun, H. J. Bae, M. G. Kim, Y. G. Chang, Q. Shen, S.-J. Kim, S. H. Kwon, W. S. Park, J. Y. Lee, and S. W. Nam, *Toxicol. Lett.*, 2013, **216**, 1–8.
76. W. Saelens, R. Cannoodt, and Y. Saeys, *Nat. Commun.*, 2018, **9**, 1090.
77. R. C. Taylor, G. Acquaah-Mensah, M. Singhal, D. Malhotra, and S. Biswal, *PLoS Comput. Biol.*, 2008, **4**, e1000166.
78. Y. Deng, D. R. Johnson, X. Guan, C. Y. Ang, J. Ai, and E. J. Perkins, *BMC Syst. Biol.*, 2010, **4**, 153.
79. G. Csárdi, Z. Kutalik, and S. Bergmann, *Bioinformatics*, 2010, **26**, 1376–1377.
80. P. Langfelder and S. Horvath, *BMC Bioinformatics*, 2008, **9**, 559.
81. Y. Ye and A. Godzik, *Genome Res.*, 2004, **14**, 343–353.
82. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, *Science*, 2002, **297**, 1551–1555.
83. P. Langfelder, R. Luo, M. C. Oldham, and S. Horvath, *PLoS Comput. Biol.*, 2011, **7**, e1001057.
84. C. Clarke, P. Doolan, N. Barron, P. Meleady, F. O'Sullivan, P. Gammell, M. Melville, M. Leonard, and M. Clynes, *J. Biotechnol.*, 2011, **155**, 350–359.
85. A. P. Presson, E. M. Sobel, J. C. Papp, C. J. Suarez, T. Whistler, M. S. Rajeevan, S. D. Vernon, and S. Horvath, *BMC Syst. Biol.*, 2008, **2**, 95.
86. B. Li, L. C. Tsoi, W. R. Swindell, J. E. Gudjonsson, T. Tejasvi, A. Johnston, J. Ding, P. E. Stuart, X. Xing, J. J. Kochkodan, J. J. Voorhees, H. M. Kang, R. P. Nair, G. R. Abecasis, and J. T. Elder, *J. Invest. Dermatol.*, 2014, **134**, 1828–1838.
87. J. A. Miller, M. C. Oldham, and D. H. Geschwind, *J. Neurosci.*, 2008, **28**, 1410–1420.
88. K. J. Conn, M. D. Ullman, M. J. Larned, P. B. Eisenhauer, R. E. Fine, and J. M. Wells, *Neurochem. Res.*, 2003.
89. I. S. Kim, D.-K. Choi, and J. H. Do, *BioChip J.*, 2013, **7**, 247–257.
90. J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, *Bioinformatics*, 2006, **22**, 507–508.
91. P. Okunieff, Y. Chen, D. J. Maguire, and A. K. Huser, *Cancer Metastasis Rev.*, 2008, **27**, 363–374.
92. D. P. Stiehl, E. Tritto, S.-D. Chibout, A. Cordier, and P. Moulin, *ILAR J.*, 2017, **58**, 69–79.
93. J. A. Rininger, V. A. DiPippo, and B. E. Gould-Rothberg, *Drug Discov. Today*, 2000, **5**, 560–568.
94. R. Sellamuthu, C. Umbright, S. Li, M. Kashon, and P. Joseph, *Inhal Toxicol*, 2011, **23**, 927–937.
95. Y. Yang, Y. Xing, C. Liang, L. Hu, F. Xu, and Q. Mei, *Tumour Biol.*, 2016, **37**, 6709–6718.
96. S. M. Bell, M. M. Angrish, C. E. Wood, and S. W. Edwards, *Toxicol. Sci.*, 2016, **150**, 510–520.
97. M. D. M. AbdulHameed, G. J. Tawa, K. Kumar, D. L. Ippolito, J. A. Lewis, J. D. Stallings, and A. Wallqvist, *PLoS One*, 2014, **9**, e112193.
98. J. A. Te, M. D. M. AbdulHameed, and A. Wallqvist, *J Appl Toxicol*, 2016, **36**, 1137–1149.
99. G. J. Tawa, M. D. M. AbdulHameed, X. Yu, K. Kumar, D. L. Ippolito, J. A. Lewis, J. D. Stallings, and A. Wallqvist, *PLoS One*, 2014, **9**, e107230.
100. M. D. M. AbdulHameed, D. L. Ippolito, J. D. Stallings, and A. Wallqvist, *BMC Genomics*,

2016, **17**, 790.
101. Y. Guo and Y. Xing, *Life Sci.*, 2016, **151**, 339–347.
102. J. J. Sutherland, Y. W. Webster, J. A. Willy, G. H. Searfoss, K. M. Goldstein, A. R. Irizarry, D. G. Hall, and J. L. Stevens, *Pharmacogenomics J*, 2018, **18**, 377–390.
103. U. G. Sauer, L. Deferme, L. Gribaldo, J. Hackermüller, T. Tralau, B. van Ravenzwaay, C. Yauk, A. Poole, W. Tong, and T. W. Gant, *Regul Toxicol Pharmacol*, 2017, **91 Suppl 1**, S14–S26.
104. A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, and A. R. Leach, *Nucleic Acids Res.*, 2017, **45**, D945–D954.
105. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, *Nucleic Acids Res.*, 2018, **46**, D1074–D1082.
106. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, *Nucleic Acids Res.*, 2006, **34**, D668-72.
107. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, in *Data Analysis, Machine Learning and Applications*, eds. C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 319–326.
108. Greg Langdrum, *RDKit: Open-source cheminformatics*, 2019.
109. A. Kassambara and F. Mundt, 2017.
110. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018.
111. L. H. Mervin, Q. Cao, I. P. Barrett, M. A. Firth, D. Murray, L. McWilliams, M. Haddrick, M. Wigglesworth, O. Engkvist, and A. Bender, *ACS Chem. Biol.*, 2016, **11**, 3007–3023.
112. L. Mervin, 2019.
113. L. H. Mervin, K. C. Bulusu, L. Kalash, A. M. Afzal, F. Svensson, M. A. Firth, I. Barrett, O. Engkvist, and A. Bender, *Bioinformatics*, 2018, **34**, 72–79.
114. B. Ganter, S. Tugendreich, C. I. Pearson, E. Ayanoglu, S. Baumhueter, K. A. Bostian, L. Brady, L. J. Browne, J. T. Calvin, G.-J. Day, N. Breckenridge, S. Dunlea, B. P. Eynon, L. M. Furness, J. Ferng, M. R. Fielden, S. Y. Fujimoto, L. Gong, C. Hu, R. Idury, M. S. B. Judo, K. L. Kolaja, M. D. Lee, C. McSorley, J. M. Minor, R. V. Nair, G. Natsoulis, P. Nguyen, S. M. Nicholson, H. Pham, A. H. Roter, D. Sun, S. Tan, S. Thode, A. M. Tolley, A. Vladimirova, J. Yang, Z. Zhou, and K. Jarnagin, *J. Biotechnol.*, 2005, **119**, 219–244.
115. T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, *Nucleic Acids Res.*, 2013, **41**, D991-5.
116. P. B. Dallas, N. G. Gottardo, M. J. Firth, A. H. Beesley, K. Hoffmann, P. A. Terry, J. R. Freitas, J. M. Boag, A. J. Cummings, and U. R. Kees, *BMC Genomics*, 2005, **6**, 59.
117. L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, *Bioinformatics*, 2004, **20**, 307–315.
118. H. Pages, M. Carlson, S. Falcon, and N. Li, *AnnotationDbi: Annotation Database Interface*, Bioconductor, 2018.
119. K. Fundel, R. Küffner, T. Aigner, and R. Zimmer, *Bioinform. Biol. Insights*, 2008, **2**, 291–305.
120. DrugMatrix database, http://ftp//anonftp.niehs.nih.gov/ntp-cebs/datatype/Drug_Matrix/annotation, [data accessed 01/09/2019], .
121. Open TG-GATEs database, https://dbarchive.biosciencedbc.jp/en/open-tggates/download.html, [date accessed 01/07/2018], .

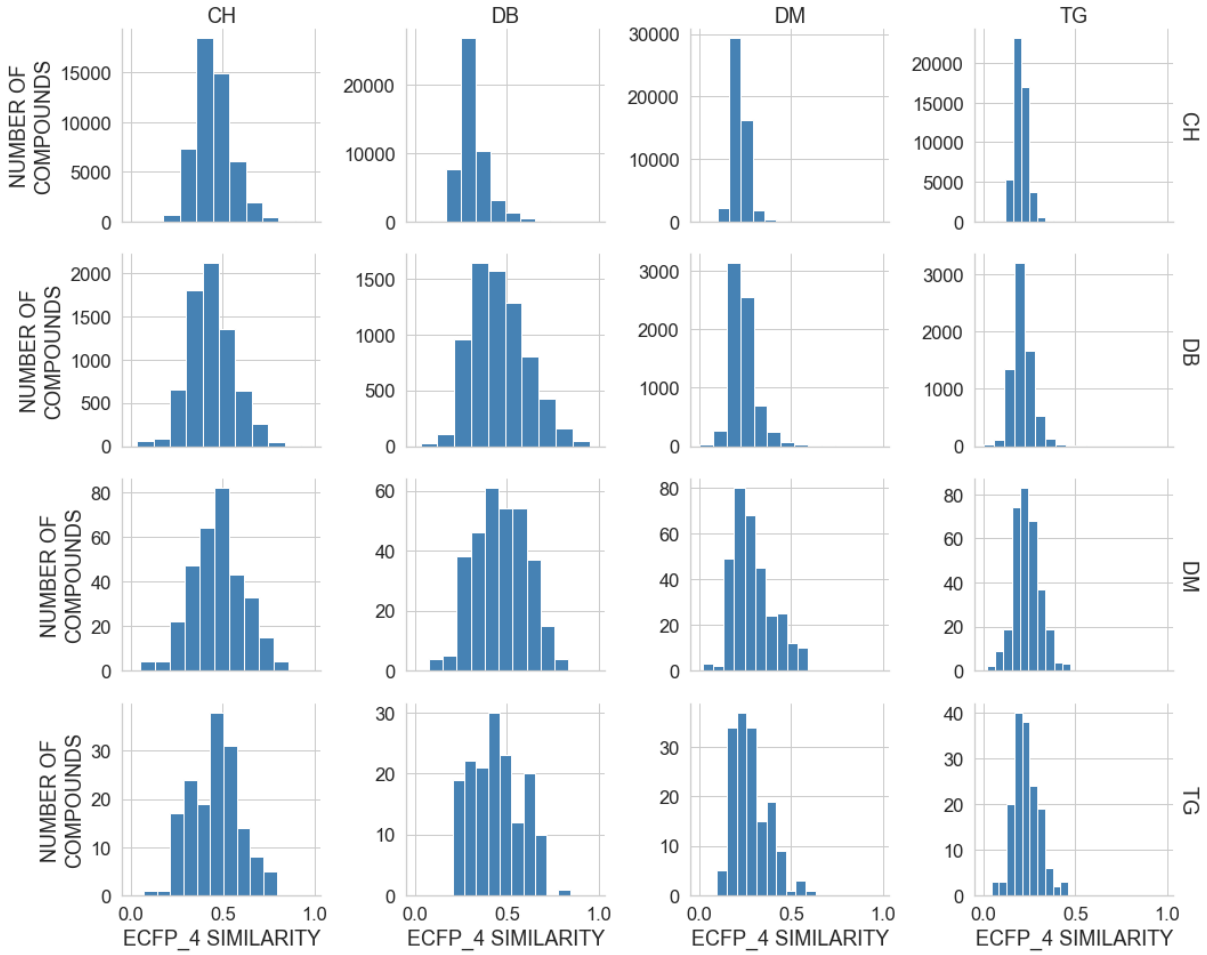
122. F. Hammann, V. Schöning, and J. Drewe, *J Appl Toxicol*, 2019, **39**, 412–419.
123. B. Chen, P. Greenside, H. Paik, M. Sirota, D. Hadley, and A. J. Butte, *CPT Pharmacometrics Syst. Pharmacol.*, 2015, **4**, 576–584.
124. A. K. Abbas, *Cellular and molecular immunology*, Saunders, Philadelphia, 5th ed., 2006.
125. S. Shetty, P. F. Lalor, and D. H. Adams, *Toxicology*, 2008, **254**, 136–146.
126. K. M. Shabana, K. A. Abdul Nazeer, M. Pradhan, and M. Palakal, *BMC Bioinformatics*, 2015, **16 Suppl 17**, S5.
127. I. Nassiri and M. N. McCall, *Nucleic Acids Res.*, 2018, **46**, e116.
128. R. R. Maronpot, *Liver, Hepatocyte - Glycogen Accumulation and Depletion National Toxicology Program Nonneoplastic Lesion Atlas*, 2014.
129. M. Krishna, *Clin Liver Dis (Hoboken)*, 2017, **10**, 53–56.
130. M. R. Fielden, R. Brennan, and J. Gollub, *Toxicol. Sci.*, 2007, **99**, 90–100.
131. M. R. Fielden, A. Adai, R. T. Dunn, A. Olaharski, G. Searfoss, J. Sina, J. Aubrecht, E. Boitier, P. Nioi, S. Auerbach, D. Jacobson-Kram, N. Raghavan, Y. Yang, A. Kincaid, J. Sherlock, S.-J. Chen, B. Car, and Predictive Safety Testing Consortium, Carcinogenicity Working Group, *Toxicol. Sci.*, 2011, **124**, 54–74.
132. M. Kanki, M. Gi, M. Fujioka, and H. Wanibuchi, *J Toxicol Sci*, 2016, **41**, 281–292.
133. J. Kim and M. Shin, *ijms*, 2017, **18**, 755.
134. J. Kim and M. Shin, *BMC Bioinformatics*, 2014, **15 Suppl 16**, S2.
135. Z. Gu, R. Eils, and M. Schlesner, *Bioinformatics*, 2016, **32**, 2847–2849.
136. M. S. Schröder, D. Gusenleitner, J. Quackenbush, A. C. Culhane, and B. Haibe-Kains, *Bioinformatics*, 2013, **29**, 666–668.
137. D. Greene, S. Richardson, and E. Turro, *Bioinformatics*, 2017, **33**, 1104–1106.
138. L. Garcia-Alonso, F. Iorio, A. Matchan, N. Fonseca, P. Jaaks, G. Peat, M. Pignatelli, F. Falcone, C. H. Benes, I. Dunham, G. Bignell, S. S. McDade, M. J. Garnett, and J. Saez-Rodriguez, *Cancer Res.*, 2018, **78**, 769–780.
139. A. Liu, P. Trairatphisan, E. Gjerga, A. Didangelos, J. Barratt, and J. Saez-Rodriguez, *BioRxiv*, 2019.
140. D. Gusenleitner, S. S. Auerbach, T. Melia, H. F. Gómez, D. H. Sherr, and S. Monti, *PLoS One*, 2014, **9**, e102579.
141. K. Kim, D. Ryu, P. Dongiovanni, L. Ozcan, S. Nayak, B. Ueberheide, L. Valenti, J. Auwerx, and U. B. Pajvani, *Gastroenterology*, 2017, **153**, 1568–1580.e10.
142. E. S. Jin, M. H. Lee, R. E. Murphy, and C. R. Malloy, *Am. J. Physiol. Endocrinol. Metab.*, 2018, **314**, E543–E551.
143. T. Jensen, M. F. Abdelmalek, S. Sullivan, K. J. Nadeau, M. Green, C. Roncal, T. Nakagawa, M. Kuwabara, Y. Sato, D.-H. Kang, D. R. Tolan, L. G. Sanchez-Lozada, H. R. Rosen, M. A. Lanaspa, A. M. Diehl, and R. J. Johnson, *J. Hepatol.*, 2018, **68**, 1063–1075.
144. M. Saffran, *Trends Endocrinol. Metab.*, 1994, **5**, 354–355.
145. C. Tran, *Nutrients*, 2017, **9**, 356.
146. T. Luedde and R. F. Schwabe, *Nat. Rev. Gastroenterol. Hepatol.*, 2011, **8**, 108–118.
147. A. P. Hall, C. R. Elcombe, J. R. Foster, T. Harada, W. Kaufmann, A. Knippel, K. Küttler, D. E. Malarkey, R. R. Maronpot, A. Nishikawa, T. Nolte, A. Schulte, V. Strauss, and M. J. York, *Toxicol. Pathol.*, 2012, **40**, 971–994.
148. R. F. Crampton, T. J. Gray, P. Grasso, and D. V. Parke, *Toxicology*, 1977, **7**, 307–326.
149. P. Grasso and R. H. Hinton, *Mutat. Res.*, 1991, **248**, 271–290.
150. R. C. Cattley and J. A. Popp, *Cancer Res.*, 1989, **49**, 3246–3251.
151. P. Grasso, M. G. Wright, S. D. Gangolli, and R. J. Hendy, *Food Cosmet Toxicol*, 1974, **12**, 341–350.
152. E. Ulusoy and B. Eren, *Clin. Med. Pathol.*, 2008, **1**, 69–75.
153. M. M. Adeva-Andany, M. González-Lucán, C. Donapetry-García, C. Fernández-Fernández,

and E. Ameneiros-Rodríguez, *BBA Clin.*, 2016, **5**, 85–100.
154. A. V. Shubin, I. V. Demidyuk, A. A. Komissarov, L. M. Rafieva, and S. V. Kostrov, *Oncotarget*, 2016, **7**, 55863–55889.
155. P. M. Dijkman and A. Watts, *Biochim. Biophys. Acta*, 2015, **1848**, 2889–2897.
156. T. Luedde, N. Kaplowitz, and R. F. Schwabe, *Gastroenterology*, 2014, **147**, 765–783.e4.
157. L. W. Lamps, *Arch. Pathol. Lab. Med.*, 2015, **139**, 867–875.
158. S. S. Bhardwaj, R. Saxena, and P. Y. Kwo, *Curr. Gastroenterol. Rep.*, 2009, **11**, 42–49.
159. D. R. Gaya, D. Thorburn, K. A. Oien, A. J. Morris, and A. J. Stanley, *J. Clin. Pathol.*, 2003, **56**, 850–853.
160. S. A. Bustin, *J. Mol. Endocrinol.*, 2000, **25**, 169–193.
161. F. M. Casares, *Med. Sci. Monit. Basic Res.*, 2016, **22**, 45–52.
162. F. Saito, in *Alternatives to animal testing: proceedings of asian congress 2016*, eds. H. Kojima, T. Seidle, and H. Spielmann, Springer Singapore, Singapore, 2019, pp. 91–104.
163. Z. Liu, B. Delavan, R. Roberts, and W. Tong, *Front. Genet.*, 2018, **9**, 74.
164. A. Mueller, J. O'Rourke, J. Grimm, K. Guillemin, M. F. Dixon, A. Lee, and S. Falkow, *Proc. Natl. Acad. Sci. USA*, 2003, **100**, 1292–1297.
165. Q. Liu, J. M. Spitsbergen, R. Cariou, C.-Y. Huang, N. Jiang, G. Goetz, R. J. Hutz, P. J. Tonellato, and M. J. Carvan, *PLoS One*, 2014, **9**, e100910.
166. H. A. Rueda-Zárate, I. Imaz-Rosshandler, R. A. Cárdenas-Ovando, J. E. Castillo-Fernández, J. Noguez-Monroy, and C. Rangel-Escareño, *PLoS One*, 2017, **12**, e0176284.
167. P. Langfelder and S. Horvath, *J Stat Softw*, 2012, **46**.
168. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
169. J. H. Morris, L. Apeltsin, A. M. Newman, J. Baumbach, T. Wittkop, G. Su, G. D. Bader, and T. E. Ferrin, *BMC Bioinformatics*, 2011, **12**, 436.
170. M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan, *Nucleic Acids Res.*, 2016, **44**, W90-7.
171. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995, **57**, 289–300.
172. G. R. Warnes, B. Bolker, L. Bonebakker, and R. Gentleman, .
173. T. W. H Backman and T. Girke, *BMC Bioinformatics*, 2016, **17**, 388.
174. S. G. Park, P. Schimmel, and S. Kim, *Proc. Natl. Acad. Sci. USA*, 2008, **105**, 11043–11049.
175. S. G. Jantzen, B. J. Sutherland, D. R. Minkley, and B. F. Koop, *BMC Res. Notes*, 2011, **4**, 267.
176. S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron, *Bioinformatics*, 2007, **23**, 3024–3031.
177. L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T.-M. Chu, F. M. Goodsaid, L. Pusztai, J. D. Shaughnessy, A. Oberthuer, R. S. Thomas, R. S. Paules, M. Fielden, B. Barlogie, W. Chen, P. Du, M. Fischer, C. Furlanello, B. D. Gallas, X. Ge, D. B. Megherbi, W. F. Symmans, M. D. Wang, J. Zhang, H. Bitter, B. Brors, P. R. Bushel, M. Bylesjo, M. Chen, J. Cheng, J. Cheng, J. Chou, T. S. Davison, M. Delorenzi, Y. Deng, V. Devanarayan, D. J. Dix, J. Dopazo, K. C. Dorff, F. Elloumi, J. Fan, S. Fan, X. Fan, H. Fang, N. Gonzaludo, K. R. Hess, H. Hong, J. Huan, R. A. Irizarry, R. Judson, D. Juraeva, S. Lababidi, C. G. Lambert, L. Li, Y. Li, Z. Li, S. M. Lin, G. Liu, E. K. Lobenhofer, J. Luo, W. Luo, M. N. McCall, Y. Nikolsky, G. A. Pennello, R. G. Perkins, R. Philip, V. Popovici, N. D. Price, F. Qian, A. Scherer, T. Shi, W. Shi, J. Sung, D. Thierry-Mieg, J. Thierry-Mieg, V. Thodima, J. Trygg, L. Vishnuvajjala, S. J. Wang, J. Wu, Y. Wu, Q. Xie, W. A. Yousef, L. Zhang, X. Zhang, S. Zhong, Y. Zhou, S. Zhu, D. Arasappan, W. Bao, A. B. Lucas, F. Berthold, R. J. Brennan, A. Buness, J. G. Catalano, C. Chang, R. Chen, Y. Cheng, J. Cui, W.

Czika, F. Demichelis, X. Deng, D. Dosymbekov, R. Eils, Y. Feng, J. Fostel, S. Fulmer-Smentek, J. C. Fuscoe, L. Gatto, W. Ge, D. R. Goldstein, L. Guo, D. N. Halbert, J. Han, S. C. Harris, C. Hatzis, D. Herman, J. Huang, R. V. Jensen, R. Jiang, C. D. Johnson, G. Jurman, Y. Kahlert, S. A. Khuder, M. Kohl, J. Li, L. Li, M. Li, Q.-Z. Li, S. Li, Z. Li, J. Liu, Y. Liu, Z. Liu, L. Meng, M. Madera, F. Martinez-Murillo, I. Medina, J. Meehan, K. Miclaus, R. A. Moffitt, D. Montaner, P. Mukherjee, G. J. Mulligan, P. Neville, T. Nikolskaya, B. Ning, G. P. Page, J. Parker, R. M. Parry, X. Peng, R. L. Peterson, J. H. Phan, B. Quanz, Y. Ren, S. Riccadonna, A. H. Roter, F. W. Samuelson, M. M. Schumacher, J. D. Shambaugh, Q. Shi, R. Shippy, S. Si, A. Smalter, C. Sotiriou, M. Soukup, F. Staedtler, G. Steiner, T. H. Stokes, Q. Sun, P.-Y. Tan, R. Tang, Z. Tezak, B. Thorn, M. Tsyganova, Y. Turpaz, S. C. Vega, R. Visintainer, J. von Frese, C. Wang, E. Wang, J. Wang, W. Wang, F. Westermann, J. C. Willey, M. Woods, S. Wu, N. Xiao, J. Xu, L. Xu, L. Yang, X. Zeng, J. Zhang, L. Zhang, M. Zhang, C. Zhao, R. K. Puri, U. Scherf, and et al., *Nat. Biotechnol.*, 2010, **28**, 827–838.

178. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-D613. doi:10.1093/nar/gky1131

# 7. Supplementary Information




*Supplementary Figure 1 distribution of compound similarities from ChEMBL (CH), DrugBank (DB), DrugMatrix (DM) and Open TG-GATEs (TG). All follow an approximate normal distribution.*

*Supplementary Table 1 Full mapping between DrugMatrix and HPATH histopathology terms*

| DrugMatrix Term | HPATH term | Term ID |
|---|---|---|
| HEART_ENDOCARDIUM, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate | MC_0000850 |
| HEART_EPICARDIUM, CELLULAR INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| HEART_EPICARDIUM, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate | MC_0000850 |
| HEART_EPICARDIUM, HYPERPLASIA | hyperplasia\mesothelial\epicardium or pericardium | MC_0000467 |
| HEART_INTERVENTRICULAR SEPTUM, CELLULAR INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| HEART_INTERVENTRICULAR SEPTUM, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate | MC_0000850 |
| HEART_INTERVENTRICULAR SEPTUM, FIBROSIS | lenticular fibrosis | MC_0000402 |
| HEART_INTERVENTRICULAR SEPTUM, MYOCYTE, DEGENERATION | muscle degeneration/necrosis | MC_2000691 |
| HEART_INTRAMYOCARDIAL ARTERIES, CELLULAR INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| HEART_INTRAMYOCARDIAL ARTERIES, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate/fibrosis\myocardium | MC_0000850 |

| | | |
|---|---|---|
| HEART_INTRAMYOCARDIAL ARTERIES, DEGENERATION | cardiomyocyte degeneration | MC_2000711 |
| HEART_INTRAMYOCARDIAL ARTERIES, NECROSIS, FIBRINOID | myocardium fibrosis | MC_0000935 |
| HEART_INTRAMYOCARDIAL ARTERIES, PERIVASCULAR EDEMA | myocardium edema | MC_2000713 |
| HEART_INTRAMYOCARDIAL ARTERIES, PERIVASCULAR FIBROSIS | myocardium fibrosis | MC_0000935 |
| HEART_INTRAMYOCARDIAL ARTERIES, TUNICA MEDIA, HYPERTROPHY | cardiomyocyte hypertrophy | MC_0000468 |
| HEART_LEFT ATRIUM, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate/fibrosis myocardium | MC_0000850 |
| HEART_LEFT ATRIUM, FIBROSIS | myocardium fibrosis | MC_0000935 |
| HEART_LEFT ATRIUM, MYOCYTE, DEGENERATION | cardiomyocyte degeneration | MC_2000711 |
| HEART_LEFT VENTRICLE, CELLULAR INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| HEART_LEFT VENTRICLE, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate/fibrosis myocardium | MC_0000850 |
| HEART_LEFT VENTRICLE, FIBROSIS | myocardium fibrosis | MC_0000935 |
| HEART_LEFT VENTRICLE, MYOCYTE, DEGENERATION | cardiomyocyte degeneration | MC_2000711 |
| HEART_PAPILLARY MUSCLE, CELLULAR INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| HEART_PAPILLARY MUSCLE, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate/fibrosis myocardium | MC_0000850 |
| HEART_PAPILLARY MUSCLE, MYOCYTE, DEGENERATION | cardiomyocyte degeneration | MC_2000711 |
| HEART_RIGHT ATRIUM, CELLULAR INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| HEART_RIGHT ATRIUM, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate/fibrosis myocardium | MC_0000850 |
| HEART_RIGHT ATRIUM, MYOCYTE, DEGENERATION | cardiomyocyte degeneration | MC_2000711 |
| HEART_RIGHT VENTRICLE, CELLULAR INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| HEART_RIGHT VENTRICLE, CELLULAR INFILTRATE, MONONUCLEAR CELL | mononuclear cell infiltrate/fibrosis myocardium | MC_0000850 |
| HEART_RIGHT VENTRICLE, FIBROSIS | myocardium fibrosis | MC_0000935 |
| HEART_RIGHT VENTRICLE, MINERALIZATION | mineralization\cardiomyocyte or myocardium | MC_2000715 |
| HEART_RIGHT VENTRICLE, MYOCYTE, DEGENERATION | cardiomyocyte degeneration | MC_2000711 |
| HEART_RIGHT VENTRICLE, MYOCYTE, LIPID ACCUMULATION | cardiomyocyte vacuolation | MC_0000476 |
| INTESTINE_EPITHELIAL ATROPHY | mucosa atrophy | MC_0000325 |
| INTESTINE_VILLUS TIP FUSION | Villus fusing | MC_0000356 |
| KIDNEY_AUTOLYSIS | autolysis | MC_2000081 |
| KIDNEY_CORTEX, CYST(S) | cyst | MC_2000164 |
| KIDNEY_CORTEX, MINERALIZATION | mineralization | MC_2000168 |
| KIDNEY_CORTEX, TUBULE, DEGENERATION | tubule degeneration | MC_0000867 |
| KIDNEY_CORTEX, TUBULE, DILATATION | tubule dilation | MC_0000483 |
| KIDNEY_CORTEX, TUBULE, INCREASED MITOSES | increased mitoses | MC_0000407 |
| KIDNEY_CORTEX, TUBULE, NECROSIS | renal tubules necrosis | MC_0000836 |
| KIDNEY_CORTEX, TUBULE, VACUOLATION | tubular vacuolation | MC_0000872 |
| KIDNEY_INFILTRATIVE CELL, POLYMORPHONUCLEAR CELL | granulocytic inflammatory cell infiltration | MC_2000145 |
| KIDNEY_INTERSTITIUM, FIBROSIS | interstitial fibrosis | MC_0000934 |

| | | |
|---|---|---|
| KIDNEY_INTERSTITIUM, INFLAMMATION, CHRONIC | lymphohistioplasmacytic inflammation | MC_2000159 |
| KIDNEY_MEDULLA, CYST(S) | cyst | MC_2000164 |
| KIDNEY_MEDULLA, MINERALIZATION | mineralization | MC_2000168 |
| KIDNEY_MEDULLA, TUBULE, DILATATION | tubule dilation | MC_0000483 |
| KIDNEY_NEPHROPATHY | nephropathy | MC_0000857 |
| KIDNEY_PAPILLA, CYST(S) | cystic/papillary hyperplasia | MC_0000648 |
| KIDNEY_PAPILLA, MINERALIZATION | mineralization | MC_2000168 |
| KIDNEY_PAPILLA, NECROSIS | papillary necrosis | MC_0000835 |
| KIDNEY_PAPILLA, TUBULE, REGENERATION | tubule regeneration | MC_0000988 |
| KIDNEY_PELVIS, DILATATION | pelvis dilation | MC_0000482 |
| KIDNEY_PELVIS, INFLAMMATION, SUBACUTE | lymphocytic inflammation | MC_0000011 |
| KIDNEY_PELVIS, UROTHELIAL HYPERPLASIA | urothelium hyperplasia | MC_0000822 |
| KIDNEY_PERIVASCULAR EDEMA | pericascular inflammatory cell infiltrate | MC_2000165 |
| KIDNEY_TUBULE, CAST, GRANULAR | granular casts | MC_2000487 |
| KIDNEY_TUBULE, CAST, PROTEINACEOUS | casts | MC_0000502 |
| KIDNEY_TUBULE, HYALINE DROPLETS | hyaline droplets accumulation | MC_0000503 |
| KIDNEY_TUBULE, REGENERATION | tubule regeneration | MC_0000988 |
| LIVER_AUTOLYSIS | autolysis | MC_2000081 |
| LIVER_BILE DUCT DILATATION | ductal dilation | MC_0000365 |
| LIVER_BILE DUCT HYPERPLASIA | bile duct hyperplasia | MC_0000541 |
| LIVER_BILE DUCT, NECROSIS, ONCOCYTIC | squamous epithelium necrosis | MC_0000839 |
| LIVER_CAPSULE, HEMORRHAGE | hemorrhage | MC_0000126 |
| LIVER_CAPSULE, INFLAMMATORY CELL INFILTRATE, LYMPHOID | lymphocytic inflammatory cell infiltration | MC_2000151 |
| LIVER_CAPSULE, INFLAMMATORY CELL INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| LIVER_CAPSULE, INFLAMMATORY CELL INFILTRATE, NEUTROPHILIC | neutrophil infiltration | MC_2000147 |
| LIVER_CAPSULE, MESOTHELIAL CELL, HYPERPLASIA | mesothelium hyperplasia | MC_0000199 |
| LIVER_CAPSULE, THROMBUS | thrombosis | MC_2000400 |
| LIVER_CENTRILOBULAR FIBROSIS | regeneration fibrosis | MC_0000528 |
| LIVER_CENTRILOBULAR, INFLAMMATORY CELL INFILTRATE, LYMPHOID | lymphocytic inflammatory cell infiltration | MC_2000151 |
| LIVER_CENTRILOBULAR, INFLAMMATORY CELL INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| LIVER_CENTRILOBULAR, INFLAMMATORY CELL INFILTRATE, NEUTROPHILIC | neutrophil infiltration | MC_2000147 |
| LIVER_CLEAR CELL FOCUS | clear cell cellular alteration | MC_0000538 |
| LIVER_HEPATOCYTE, CENTRILOBULAR, ATROPHY | hepatocyte atrophy | MC_0000532 |
| LIVER_HEPATOCYTE, CENTRILOBULAR, CYTOPLASM, EOSINOPHILIA | cytoplasmic alteration | MC_2000255 |
| LIVER_HEPATOCYTE, CENTRILOBULAR, DEGENERATION | degeneration | MC_0000101 |
| LIVER_HEPATOCYTE, CENTRILOBULAR, GLYCOGEN ACCUMULATION | glycogen accumulation | MC_0000551 |
| LIVER_HEPATOCYTE, CENTRILOBULAR, HYPERTROPHY | hepatocyte hypertrophy | MC_0000533 |

| | | |
|---|---|---|
| LIVER_HEPATOCYTE, CENTRILOBULAR, LIPID ACCUMULATION, MACROVESICULAR | lipidic vacuolation (fatty change) | MC_0000515 |
| LIVER_HEPATOCYTE, CENTRILOBULAR, LIPID ACCUMULATION, MICROVESICULAR | lipidic vacuolation (fatty change) | MC_0000515 |
| LIVER_HEPATOCYTE, CENTRILOBULAR, NECROSIS, APOPTOTIC | zonal centrilobular necrosis | MC_0000844 |
| LIVER_HEPATOCYTE, CENTRILOBULAR, NECROSIS, ONCOCYTIC | zonal centrilobular necrosis | MC_0000844 |
| LIVER_HEPATOCYTE, DIFFUSE, CYTOPLASM, EOSINOPHILIA | cytoplasmic alteration | MC_2000255 |
| LIVER_HEPATOCYTE, DIFFUSE, HYPERTROPHY | hepatocyte hypertrophy | MC_0000533 |
| LIVER_HEPATOCYTE, MIDZONAL, LIPID ACCUMULATION, MACROVESICULAR | lipidic vacuolation (fatty change) | MC_0000515 |
| LIVER_HEPATOCYTE, MIDZONAL, LIPID ACCUMULATION, MICROVESICULAR | lipidic vacuolation (fatty change) | MC_0000515 |
| LIVER_HEPATOCYTE, MIDZONAL, NECROSIS, APOPTOTIC | zonal\midzonal necrosis | MC_0000846 |
| LIVER_HEPATOCYTE, MIDZONAL, NECROSIS, ONCOCYTIC | zonal\midzonal necrosis | MC_0000846 |
| LIVER_HEPATOCYTE, NONZONAL, DEGENERATION | degeneration | MC_0000101 |
| LIVER_HEPATOCYTE, NONZONAL, ERYTHROPHAGOCYTOSIS | erythrophagocytosis | MC_0000239 |
| LIVER_HEPATOCYTE, NONZONAL, GLYCOGEN ACCUMULATION | glycogen accumulation | MC_0000551 |
| LIVER_HEPATOCYTE, NONZONAL, INCREASED MITOSES | Increased mitoses | MC_0000407 |
| LIVER_HEPATOCYTE, NONZONAL, LIPID ACCUMULATION, MACROVESICULAR | lipidic vacuolation (fatty change) | MC_0000515 |
| LIVER_HEPATOCYTE, NONZONAL, LIPID ACCUMULATION, MICROVESICULAR | lipidic vacuolation (fatty change) | MC_0000515 |
| LIVER_HEPATOCYTE, NONZONAL, MINERALIZATION | mineralization | MC_2000168 |
| LIVER_HEPATOCYTE, NONZONAL, MULTINUCLEATED | hepatic karyocytomegaly and/or multinucleated hepatocytes | MC_2000291 |
| LIVER_HEPATOCYTE, NONZONAL, NECROSIS, APOPTOTIC | hepatocellular necrosis | MC_0000831 |
| LIVER_HEPATOCYTE, NONZONAL, NECROSIS, ONCOCYTIC | hepatocellular necrosis | MC_0000831 |
| LIVER_HEPATOCYTE, PERIPORTAL, CYTOPLASM, EOSINOPHILIA | eosinophil inflammatory cell infiltration | MC_2000149 |
| LIVER_HEPATOCYTE, PERIPORTAL, GLYCOGEN ACCUMULATION | glycogen accumulation | MC_0000551 |
| LIVER_HEPATOCYTE, PERIPORTAL, HYPERTROPHY | hepatocyte hypertrophy | MC_0000533 |
| LIVER_HEPATOCYTE, PERIPORTAL, LIPID ACCUMULATION, MACROVESICULAR | lipidic vacuolation (fatty change) | MC_0000515 |
| LIVER_HEPATOCYTE, PERIPORTAL, LIPID ACCUMULATION, MICROVESICULAR | lipidic vacuolation (fatty change) | MC_0000515 |
| LIVER_HEPATOCYTE, PERIPORTAL, NECROSIS, APOPTOTIC | zonal/periportal necrosis | MC_0000847 |
| LIVER_HEPATOCYTE, PERIPORTAL, NECROSIS, ONCOCYTIC | zonal/periportal necrosis | MC_0000847 |
| LIVER_HEPATOCYTE, SUBCAPSULAR, MINERALIZATION | mineralization | MC_2000168 |
| LIVER_HEPATOCYTE, SUBCAPSULAR, NECROSIS, ONCOCYTIC | hepatocellular necrosis | MC_0000831 |
| LIVER_MALIGNANT LYMPHOMA | malignant lymphoma | MC_0000130 |
| LIVER_MIDZONAL, INFLAMMATORY CELL INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| LIVER_NONZONAL, EXTRAMEDULLARY HEMATOPOIESIS | Extramedullary hematopoiesis | MC_0000082 |
| LIVER_NONZONAL, INFLAMMATORY CELL INFILTRATE, GRANULOMATOUS | granulomatous inflammation | MC_2000161 |
| LIVER_NONZONAL, INFLAMMATORY CELL INFILTRATE, LYMPHOID | lymphocytic inflammatory cell infiltration | MC_2000151 |
| LIVER_NONZONAL, INFLAMMATORY CELL INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |

| | | |
|---|---|---|
| LIVER_NUCLEAR CYTOPLASMIC CONDENSATION | tubular degeneration | MC_0000742 |
| LIVER_OVAL CELL, HYPERPLASIA | oval cell hyperplasia | MC_0000544 |
| LIVER_PERIPORTAL, EDEMA | edema | MC_2000376 |
| LIVER_PERIPORTAL, FIBROSIS | regeneration fibrosis | MC_0000528 |
| LIVER_PERIPORTAL, INFLAMMATORY CELL INFILTRATE, LYMPHOID | lymphocytic inflammatory cell infiltration | MC_2000151 |
| LIVER_PERIPORTAL, INFLAMMATORY CELL INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| LIVER_PERIPORTAL, INFLAMMATORY CELL INFILTRATE, NEUTROPHILIC | neutrophil infiltration | MC_2000147 |
| LIVER_PERIPORTAL, MINERALIZATION | mineralization | MC_2000168 |
| LIVER_SINGLE HEPATOCYTE NECROSIS | hepatocellular necrosis | MC_0000831 |
| LIVER_SUBCAPSULAR, FIBROSIS | regeneration fibrosis | MC_0000528 |
| LIVER_SUBCAPSULAR, INFLAMMATORY CELL INFILTRATE, MIXED CELL | mixed infiltration | MC_0000008 |
| LIVER_SUBCAPSULAR, MINERALIZATION | mineralization | MC_2000168 |
| MESENTERY_INTESTINE AUTOLYSIS | autolysis | MC_2000081 |
| MESENTERY_LYMPH NODE EDEMA | edema | MC_2000376 |
| MESENTERY_LYMPH NODE HEMORRHAGE | hemorrhage | MC_0000126 |
| MESENTERY_LYMPH NODE PROLIFERATION | Lymphocyte proliferation | MC_0000175 |
| MESENTERY_VASCULITIS | vascular/perivascular inflammation | MC_2000167 |
| SPLEEN_EXTRAMEDULLARY HEMATOPOIESIS INCREASED | Extramedullary hematopoiesis | MC_0000082 |
| SPLEEN_LYMPHOID DEPLETION | lymphoid depletion | MC_0000173 |
| THIGH MUSCLE_RIGHT BICEPS FEMORIS, MYOCYTE, DEGENERATION | muscle degeneration/necrosis | MC_2000691 |
| THIGH MUSCLE_RIGHT BICEPS FEMORIS, MYOCYTE, INFLAMMATION | inflammation | MC_0000010 |
| THIGH MUSCLE_RIGHT BICEPS FEMORIS, MYOCYTE, REGENERATION | regeneration | MC_0000893 |
| THIGH MUSCLE_RIGHT GASTROCNEMIUS, MYOCYTE, DEGENERATION | muscle degeneration/necrosis | MC_2000691 |
| THIGH MUSCLE_RIGHT GASTROCNEMIUS, MYOCYTE, INFLAMMATION | inflammation | MC_0000010 |
| THIGH MUSCLE_RIGHT GASTROCNEMIUS, MYOCYTE, REGENERATION | regeneration | MC_0000893 |
| THIGH MUSCLE_RIGHT SOLEUS, MYOCYTE, DEGENERATION | muscle degeneration/necrosis | MC_2000691 |
| THIGH MUSCLE_RIGHT SOLEUS, MYOCYTE, INFLAMMATION | inflammation | MC_0000010 |
| THIGH MUSCLE_RIGHT SOLEUS, MYOCYTE, REGENERATION | regeneration | MC_0000893 |

*Supplementary Table 2: The compound-dose instances that make up the histopathology signatures from Open TG-GATEs.*

| Compound | Dose (mg/kg) | Histopathology signature |
|---|---|---|
| benziodarone | 300 | Increased mitosis, hypertrophy |
| bendazac | 300 | |
| phenobarbital | 100 | |
| ciprofloxacin | 1000 | Microgranuloma |
| simvastatin | 120 | |
| simvastatin | 400 | |
| ajmaline | 300 | |
| dantrolene | 25 | |
| dantrolene | 75 | |
| dantrolene | 250 | |
| acarbose | 100 | |
| acarbose | 300 | |
| etoposide | 3 | |
| nimesulide | 10 | |
| nimesulide | 30 | |
| ethanol | 1200 | |
| cyclophosphamide | 1.5 | |
| cyclophosphamide | 5 | |
| cyclophosphamide | 15 | |
| desmopressin acetate | 200 | |
| methimazole | 100 | hypertrophy |
| flutamide | 150 | |
| chlormezanone | 500 | |
| imipramine | 100 | |
| hydroxyzine | 100 | |
| diltiazem | 240 | |
| diltiazem | 800 | |
| chlorpropamide | 300 | |
| bendazac | 100 | |
| phenacetin | 1000 | |
| ticlopidine | 100 | |
| ticlopidine | 300 | |
| nimesulide | 100 | |
| phenobarbital | 30 | |
| omeprazole | 1000 | |

| | | |
|---|---|---|
| benzbromarone | 60 | |
| benzbromarone | 200 | |
| diazepam | 250 | |
| bromobenzene | 100 | |
| bromobenzene | 300 | |
| ketoconazole | 100 | cytoplasmic vacuolization |
| disulfiram | 600 | |
| ethionamide | 100 | |
| carbon tetrachloride | 30 | fatty degeneration |
| carbon tetrachloride | 100 | fatty degeneration, cellular infiltration, hydropic degeneration |
| carbon tetrachloride | 300 | |
| acetamidofluorene | 30 | swelling |
| acetamidofluorene | 100 | |
| acetamidofluorene | 300 | |
| sulfasalazine | 1000 | |
| fluphenazine | 20 | necrosis |
| chlormezanone | 150 | |
| ethinylestradiol | 3 | |
| methyldopa | 60 | |
| tetracycline | 100 | |
| amitriptyline | 15 | |
| ranitidine | 300 | |
| enalapril | 600 | |
| simvastatin | 40 | |
| promethazine | 20 | |
| lornoxicam | 1 | |
| etoposide | 10 | |
| ethionamide | 30 | |
| omeprazole | 100 | |
| fluoxetine hydrochloride | 3 | |
| fluoxetine hydrochloride | 10 | |
| tamoxifen | 60 | eosinophilic change |
| bucetin | 1000 | |
| aspirin | 450 | |
| metformin | 300 | glycogen deposit |
| metformin | 1000 | |
| valproic acid | 45 | cellular infiltration |

| | | |
|---|---|---|
| ethionine | 25 | |
| amiodarone | 20 | |
| amiodarone | 200 | |
| allyl alcohol | 3 | |
| carbamazepine | 300 | |
| nitrofurantoin | 30 | |
| nitrofurantoin | 100 | |
| griseofulvin | 300 | increased mitosis |
| griseofulvin | 1000 | |
| colchicine | 1.5 | |
| benziodarone | 30 | |
| benziodarone | 100 | |
| bendazac | 30 | |
| hexachlorobenzene | 30 | |
| hexachlorobenzene | 100 | |
| hexachlorobenzene | 300 | |
| gemfibrozil | 30 | |
| gemfibrozil | 100 | |
| gemfibrozil | 300 | |
| phenylbutazone | 200 | |
| fenofibrate | 100 | increased mitosis, granular eosinophilic degeneration |
| fenofibrate | 1000 | |
| WY-14643 | 10 | |
| WY-14643 | 30 | |

*Supplementary Table 3 Enriched pathways from differentially expressed genes for each toxic group. Only significantly enriched pathways are shown (FDR p value > 0.05)*

| TermID | Number DEGs in pathway | All genes in pathway | p value | FDR corrected p value | compound and dose | toxic group |
|---|---|---|---|---|---|---|
| PW:0002106 desmosterolosis pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0000728 statin pharmacodynamics pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0001586 Wolman disease pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0001933 hypercholesterolemia pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0002270 zoledronate pharmacodynamics pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0002177 pamidronate pharmacodynamics pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0001624 nitrogenous bisphosphonate pharmacodynamics pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PW:0001812 mevalonic aciduria pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0002591 X-linked dominant chondrodysplasia punctata 2 pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0001650 Smith-Lemli-Opitz Syndrome pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0002347 cholesterol ester storage disease pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0002002 ibandronate pharmacodynamics pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0001838 risedronate pharmacodynamics pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0002174 alendronate pharmacodynamics pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |
| PW:0002439 congenital hemidysplasia with ichthyosiform erythroderma and limb defects pathway | 8 | 21 | 1.49E-17 | 9.94E-16 | fluvastatin 5 mg/kg | Group 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PW:0000454 cholesterol biosynthetic pathway | 8 | 23 | 3.58E-17 | 2.24E-15 | fluvastatin 5 mg/kg | Group 1 |
| PW:0000184 terpenoid biosynthetic pathway | 4 | 13 | 8.57E-09 | 5.05E-07 | fluvastatin 5 mg/kg | Group 1 |
| PW:0001152 steroid biosynthetic pathway | 4 | 14 | 1.20E-08 | 6.66E-07 | fluvastatin 5 mg/kg | Group 1 |
| PW:0000069 ketone bodies metabolic pathway | 2 | 10 | 1.68E-04 | 8.84E-03 | fluvastatin 5 mg/kg | Group 1 |
| PW:0000834 bile acid transport pathway | 3 | 62 | 2.41E-04 | 1.21E-02 | fluvastatin 5 mg/kg | Group 1 |
| GO:0006695 cholesterol biosynthetic process | 6 | 27 | 9.49E-12 | 1.18E-07 | fluvastatin 5 mg/kg | Group 1 |
| GO:0007584 response to nutrient | 6 | 126 | 1.39E-07 | 8.62E-04 | fluvastatin 5 mg/kg | Group 1 |
| GO:0008299 isoprenoid biosynthetic process | 3 | 13 | 1.95E-06 | 7.81E-03 | fluvastatin 5 mg/kg | Group 1 |
| GO:0016126 sterol biosynthetic process | 3 | 15 | 3.09E-06 | 7.81E-03 | fluvastatin 5 mg/kg | Group 1 |
| GO:0010949 negative regulation of intestinal phytosterol absorption | 2 | 2 | 3.77E-06 | 7.81E-03 | fluvastatin 5 mg/kg | Group 1 |
| GO:0045796 negative regulation of | 2 | 2 | 3.77E-06 | 7.81E-03 | fluvastatin 5 mg/kg | Group 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| intestinal cholesterol absorption | | | | | | |
| PW:0000375 phase I biotransformation pathway via cytochrome P450 | 12 | 70 | 1.03E-10 | 1.03E-07 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| PW:0000523 linoleic acid metabolic pathway | 7 | 33 | 1.95E-07 | 9.76E-05 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| PW:0000054 tryptophan metabolic pathway | 7 | 38 | 5.45E-07 | 1.82E-04 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| PW:0000141 retinol metabolic pathway | 8 | 66 | 2.31E-06 | 5.78E-04 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| PW:0000062 ascorbate and aldarate metabolic pathway | 4 | 16 | 4.67E-05 | 9.35E-03 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| PW:0001158 alpha-linoleic acid metabolic pathway | 4 | 17 | 6.04E-05 | 1.01E-02 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| PW:0000460 arachidonic acid metabolic pathway | 6 | 62 | 1.60E-04 | 2.28E-02 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| GO:0009617 response to bacterium | 13 | 129 | 1.44E-08 | 1.79E-04 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| GO:0060700 regulation of ribonuclease activity | 4 | 5 | 1.44E-07 | 8.94E-04 | fluocinolone acetonide 2.5 mg/kg | Group 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| GO:0055114 oxidation-reduction process | 21 | 454 | 5.63E-07 | 2.33E-03 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| GO:0006805 xenobiotic metabolic process | 7 | 49 | 3.28E-06 | 1.02E-02 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| GO:0006082 organic acid metabolic process | 6 | 35 | 5.66E-06 | 1.41E-02 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| GO:0051412 response to corticosterone | 6 | 37 | 7.93E-06 | 1.64E-02 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| GO:0017144 drug metabolic process | 5 | 23 | 1.04E-05 | 1.84E-02 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| GO:0051085 chaperone cofactor-dependent protein refolding | 5 | 27 | 2.39E-05 | 3.71E-02 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| GO:0016064 immunoglobulin mediated immune response | 4 | 15 | 3.54E-05 | 4.89E-02 | fluocinolone acetonide 2.5 mg/kg | Group 1 |
| PW:0000040 steroid hormone biosynthetic pathway | 6 | 44 | 8.39E-07 | 8.40E-04 | ethinylestradiol 10 mg/kg | Group 1 |
| PW:0000523 linoleic acid metabolic pathway | 5 | 33 | 4.23E-06 | 2.12E-03 | ethinylestradiol 10 mg/kg | Group 1 |
| PW:0000834 bile acid transport pathway | 5 | 62 | 9.72E-05 | 3.24E-02 | ethinylestradiol 10 mg/kg | Group 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PW:0000141 retinol metabolic pathway | 5 | 66 | 1.31E-04 | 3.24E-02 | ethinylestradiol 10 mg/kg | Group 1 |
| PW:0000375 phase I biotransformation pathway via cytochrome P450 | 5 | 69 | 1.62E-04 | 3.24E-02 | ethinylestradiol 10 mg/kg | Group 1 |
| PW:0000062 ascorbate and aldarate metabolic pathway | 3 | 16 | 2.11E-04 | 3.52E-02 | ethinylestradiol 10 mg/kg | Group 1 |
| PW:0001158 alpha-linoleic acid metabolic pathway | 3 | 17 | 2.55E-04 | 3.64E-02 | ethinylestradiol 10 mg/kg | Group 1 |
| GO:0055114 oxidation-reduction process | 20 | 455 | 1.25E-10 | 1.55E-06 | ethinylestradiol 10 mg/kg | Group 1 |
| PW:0000054 tryptophan metabolic pathway | 4 | 38 | 4.35E-07 | 4.35E-04 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000062 ascorbate and aldarate metabolic pathway | 3 | 16 | 2.31E-06 | 1.15E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000188 pentose and glucuronate interconversion pathway | 3 | 19 | 3.98E-06 | 1.33E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000077 beta-alanine metabolic pathway | 3 | 26 | 1.06E-05 | 2.38E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000051 histidine metabolic pathway | 3 | 27 | 1.19E-05 | 2.38E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PW:0000064 propanoate metabolic pathway | 3 | 36 | 2.88E-05 | 4.75E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000073 lysine degradation pathway | 3 | 38 | 3.39E-05 | 4.75E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000058 fatty acid metabolic pathway | 3 | 41 | 4.27E-05 | 4.75E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0001156 glycerolipid metabolic pathway | 3 | 41 | 4.27E-05 | 4.75E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000043 pyruvate metabolic pathway | 3 | 46 | 6.05E-05 | 5.50E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000071 valine, leucine and isoleucine degradation pathway | 3 | 46 | 6.05E-05 | 5.50E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000050 arginine and proline metabolic pathway | 3 | 51 | 8.25E-05 | 6.88E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000025 glycolysis/gluconeogenesis pathway | 3 | 55 | 1.03E-04 | 7.97E-03 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000640 glycolysis pathway | 3 | 61 | 1.41E-04 | 1.01E-02 | n nitrosodiethylamine 100 mg/kg | Group 2 |
| PW:0000641 gluconeogenesis pathway | 3 | 65 | 1.70E-04 | 1.14E-02 | n nitrosodiethylamine 100 mg/kg | Group 2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PW:0000375 phase I biotransformation pathway via cytochrome P450 | 5 | 69 | 2.11E-07 | 2.11E-04 | pantoprazole 1100 mg/kg | Group 2 |
| PW:0000054 tryptophan metabolic pathway | 4 | 38 | 8.55E-07 | 4.28E-04 | pantoprazole 1100 mg/kg | Group 2 |
| PW:0000141 retinol metabolic pathway | 4 | 66 | 8.04E-06 | 2.68E-03 | pantoprazole 1100 mg/kg | Group 2 |
| PW:0000134 glutathione metabolic pathway | 3 | 47 | 1.06E-04 | 2.64E-02 | pantoprazole 1100 mg/kg | Group 2 |
| PW:0000050 arginine and proline metabolic pathway | 3 | 51 | 1.35E-04 | 2.70E-02 | pantoprazole 1100 mg/kg | Group 2 |
| PW:0000460 arachidonic acid metabolic pathway | 3 | 62 | 2.41E-04 | 4.03E-02 | pantoprazole 1100 mg/kg | Group 2 |
| GO:0055114 oxidation-reduction process | 9 | 454 | 1.26E-07 | 1.57E-03 | pantoprazole 1100 mg/kg | Group 2 |
| GO:0098869 cellular oxidant detoxification | 4 | 63 | 6.67E-06 | 2.96E-02 | pantoprazole 1100 mg/kg | Group 2 |
| GO:0042744 hydrogen peroxide catabolic process | 3 | 21 | 8.96E-06 | 2.96E-02 | pantoprazole 1100 mg/kg | Group 2 |
| GO:0046223 aflatoxin catabolic process | 2 | 3 | 1.13E-05 | 2.96E-02 | pantoprazole 1100 mg/kg | Group 2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| GO:0017144 drug metabolic process | 3 | 23 | 1.19E-05 | 2.96E-02 | pantoprazole 1100 mg/kg | Group 2 |
| PW:0000058 fatty acid metabolic pathway | 5 | 43 | 6.16E-09 | 6.17E-06 | methyl salicylate 444 mg/kg | Group 3 |
| PW:0001137 unsaturated fatty acid biosynthetic pathway | 4 | 20 | 2.41E-08 | 1.20E-05 | methyl salicylate 444 mg/kg | Group 3 |
| PW:0000071 valine, leucine and isoleucine degradation pathway | 4 | 47 | 8.61E-07 | 2.87E-04 | methyl salicylate 444 mg/kg | Group 3 |
| PW:0001147 eicosanoid signaling pathway via peroxisome proliferator-activated receptor gamma | 4 | 68 | 3.85E-06 | 9.62E-04 | methyl salicylate 444 mg/kg | Group 3 |
| GO:0006635 fatty acid beta-oxidation | 4 | 45 | 7.21E-07 | 8.96E-03 | methyl salicylate 444 mg/kg | Group 3 |
| GO:1902380 positive regulation of endoribonuclease activity | 2 | 2 | 2.48E-06 | 1.54E-02 | methyl salicylate 444 mg/kg | Group 3 |
| GO:0010124 phenylacetate catabolic process | 2 | 3 | 7.43E-06 | 3.08E-02 | methyl salicylate 444 mg/kg | Group 3 |
| GO:0070370 cellular heat acclimation | 2 | 4 | 1.48E-05 | 3.69E-02 | methyl salicylate 444 mg/kg | Group 3 |

| | | | | | | |
|---|---|---|---|---|---|---|
| GO:0070434 positive regulation of nucleotide-binding oligomerization domain containing 2 signaling pathway | 2 | 4 | 1.48E-05 | 3.69E-02 | methyl salicylate 444 mg/kg | Group 3 |
| GO:0097201 negative regulation of transcription from RNA polymerase II promoter in response to stress | 2 | 5 | 2.47E-05 | 4.39E-02 | methyl salicylate 444 mg/kg | Group 3 |
| GO:0090063 positive regulation of microtubule nucleation | 2 | 5 | 2.47E-05 | 4.39E-02 | methyl salicylate 444 mg/kg | Group 3 |
| PW:0001026 graft-versus-host disease pathway | 3 | 32 | 7.78E-06 | 4.05E-03 | clotrimazole 52 mg/kg | Group 3 |
| PW:0001025 allograft rejection pathway | 3 | 34 | 9.37E-06 | 4.05E-03 | clotrimazole 52 mg/kg | Group 3 |
| PW:0001024 autoimmune thyroiditis pathway | 3 | 37 | 1.21E-05 | 4.05E-03 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000239 type 1 diabetes mellitus pathway | 3 | 42 | 1.79E-05 | 4.47E-03 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000050 arginine and proline metabolic pathway | 3 | 51 | 3.22E-05 | 6.44E-03 | clotrimazole 52 mg/kg | Group 3 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PW:0000825 antigen processing and presentation pathway | 3 | 56 | 4.27E-05 | 7.12E-03 | clotrimazole 52 mg/kg | Group 3 |
| PW:0001037 myocarditis pathway | 3 | 62 | 5.80E-05 | 8.29E-03 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000141 retinol metabolic pathway | 3 | 66 | 6.99E-05 | 8.75E-03 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000375 phase I biotransformation pathway via cytochrome P450 | 3 | 69 | 7.99E-05 | 8.88E-03 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000062 ascorbate and aldarate metabolic pathway | 2 | 16 | 1.73E-04 | 1.74E-02 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000188 pentose and glucuronate interconversion pathway | 2 | 19 | 2.46E-04 | 2.24E-02 | clotrimazole 52 mg/kg | Group 3 |
| PW:0001145 phagocytosis pathway | 3 | 120 | 4.12E-04 | 3.43E-02 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000077 beta-alanine metabolic pathway | 2 | 26 | 4.66E-04 | 3.59E-02 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000051 histidine metabolic pathway | 2 | 27 | 5.03E-04 | 3.60E-02 | clotrimazole 52 mg/kg | Group 3 |
| PW:0000834 bile acid transport pathway | 5 | 62 | 6.93E-05 | 3.86E-02 | diethylstilbestrol 2 8 mg/kg | Group 3 |
| PW:0000523 linoleic acid metabolic pathway | 4 | 33 | 7.70E-05 | 3.86E-02 | diethylstilbestrol 2 8 mg/kg | Group 3 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PW:0000062 ascorbate and aldarate metabolic pathway | 3 | 16 | 1.71E-04 | 4.83E-02 | diethylstilbestrol 2 8 mg/kg | Group 3 |
| PW:0001158 alpha-linoleic acid metabolic pathway | 3 | 17 | 2.07E-04 | 4.83E-02 | diethylstilbestrol 2 8 mg/kg | Group 3 |
| PW:0000040 steroid hormone biosynthetic pathway | 4 | 44 | 2.41E-04 | 4.83E-02 | diethylstilbestrol 2 8 mg/kg | Group 3 |
| GO:0055114 oxidation-reduction process | 16 | 454 | 8.44E-08 | 1.05E-03 | diethylstilbestrol 2 8 mg/kg | Group 3 |
| PW:0001670 sarcosinemia pathway | 3 | 25 | 8.87E-07 | 2.22E-04 | lomustine 4 2 mg/kg | Group 3 |
| PW:0002209 dimethylglycine dehydrogenase deficiency pathway | 3 | 25 | 8.87E-07 | 2.22E-04 | lomustine 4 2 mg/kg | Group 3 |
| PW:0001808 nonketotic hyperglycinemia pathway | 3 | 25 | 8.87E-07 | 2.22E-04 | lomustine 4 2 mg/kg | Group 3 |
| PW:0002210 dihydropyrimidine dehydrogenase deficiency pathway | 3 | 25 | 8.87E-07 | 2.22E-04 | lomustine 4 2 mg/kg | Group 3 |
| PW:0000047 glycine, serine and threonine metabolic pathway | 3 | 30 | 1.56E-06 | 3.13E-04 | lomustine 4 2 mg/kg | Group 3 |
| GO:0031667 response to nutrient levels | 4 | 136 | 2.95E-06 | 2.20E-02 | lomustine 4 2 mg/kg | Group 3 |

| | | | | | | |
|---|---|---|---|---|---|---|
| GO:0006564 L-serine biosynthetic process | 2 | 4 | 3.54E-06 | 2.20E-02 | lomustine 4 2 mg/kg | Group 3 |
| PW:0000375 phase I biotransformation pathway via cytochrome P450 | 5 | 69 | 7.42E-06 | 7.43E-03 | clonazepam 2500 mg/kg | Group 3 |
| PW:0001565 doxorubicin pharmacokinetics pathway | 3 | 17 | 3.82E-05 | 1.91E-02 | clonazepam 2500 mg/kg | Group 3 |
| PW:0000141 retinol metabolic pathway | 4 | 66 | 1.31E-04 | 4.38E-02 | clonazepam 2500 mg/kg | Group 3 |
| GO:0055114 oxidation-reduction process | 15 | 454 | 1.40E-10 | 1.74E-06 | clonazepam 2500 mg/kg | Group 3 |
| GO:0017144 drug metabolic process | 4 | 23 | 1.83E-06 | 1.14E-02 | clonazepam 2500 mg/kg | Group 3 |
| GO:0007568 aging | 9 | 312 | 3.26E-06 | 1.35E-02 | clonazepam 2500 mg/kg | Group 3 |
| GO:0033993 response to lipid | 3 | 37 | 3.97E-06 | 3.13E-02 | carmustine 4 mg/kg | Group 4 |
| GO:0120163 negative regulation of cold-induced thermogenesis | 3 | 40 | 5.03E-06 | 3.13E-02 | carmustine 4 mg/kg | Group 4 |
| GO:0042752 regulation of circadian rhythm | 4 | 43 | 1.42E-06 | 1.49E-02 | raloxifene 650 mg/kg | Group 4 |
| GO:0007623 circadian rhythm | 5 | 113 | 2.48E-06 | 1.49E-02 | raloxifene 650 mg/kg | Group 4 |

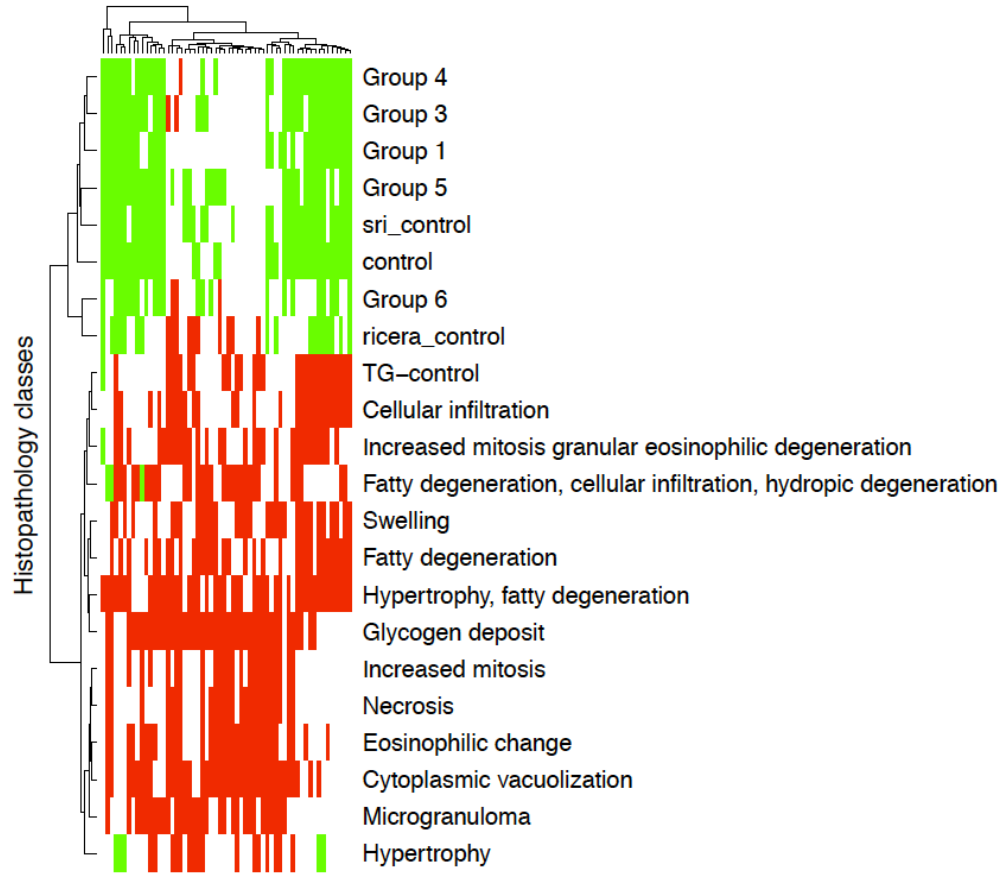
| | | | | | | |
|---|---|---|---|---|---|---|
| GO:0032922 circadian regulation of gene expression | 4 | 54 | 3.59E-06 | 1.49E-02 | raloxifene 650 mg/kg | Group 4 |
| PW:0001147 eicosanoid signaling pathway via peroxisome proliferator-activated receptor gamma | 3 | 67 | 1.81E-05 | 1.81E-02 | bis(2 ethylhexyl)phthalate 1000 mg/kg | Group 4 |
| PW:0000141 retinol metabolic pathway | 4 | 66 | 5.84E-06 | 3.49E-03 | artemisinin 2000 mg/kg | Group 4 |
| PW:0000375 phase I biotransformation pathway via cytochrome P450 | 4 | 69 | 6.98E-06 | 3.49E-03 | artemisinin 2000 mg/kg | Group 4 |
| GO:0014070 response to organic cyclic compound | 8 | 252 | 9.95E-09 | 1.24E-04 | artemisinin 2000 mg/kg | Group 4 |
| GO:0017144 drug metabolic process | 3 | 23 | 9.38E-06 | 3.82E-02 | artemisinin 2000 mg/kg | Group 4 |
| GO:0046223 aflatoxin catabolic process | 2 | 3 | 9.65E-06 | 3.82E-02 | artemisinin 2000 mg/kg | Group 4 |
| GO:0055114 oxidation-reduction process | 7 | 454 | 1.23E-05 | 3.82E-02 | artemisinin 2000 mg/kg | Group 4 |
| PW:0000007 mitogen activated protein kinase signaling pathway | 4 | 245 | 1.19E-05 | 1.19E-02 | olanzapine 23 mg/kg | Group 5 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PW:0001054 influenza A pathway | 3 | 127 | 6.34E-05 | 3.17E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:1902380 positive regulation of endoribonuclease activity | 2 | 2 | 3.86E-07 | 4.81E-03 | olanzapine 23 mg/kg | Group 5 |
| GO:0070370 cellular heat acclimation | 2 | 4 | 2.32E-06 | 8.00E-03 | olanzapine 23 mg/kg | Group 5 |
| GO:0070434 positive regulation of nucleotide-binding oligomerization domain containing 2 signaling pathway | 2 | 4 | 2.32E-06 | 8.00E-03 | olanzapine 23 mg/kg | Group 5 |
| GO:0060395 SMAD protein signal transduction | 3 | 49 | 3.60E-06 | 8.00E-03 | olanzapine 23 mg/kg | Group 5 |
| GO:0097201 negative regulation of transcription from RNA polymerase II promoter in response to stress | 2 | 5 | 3.86E-06 | 8.00E-03 | olanzapine 23 mg/kg | Group 5 |
| GO:0090063 positive regulation of microtubule nucleation | 2 | 5 | 3.86E-06 | 8.00E-03 | olanzapine 23 mg/kg | Group 5 |
| GO:0009408 response to heat | 3 | 70 | 1.06E-05 | 1.73E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0090084 negative regulation of | 2 | 9 | 1.39E-05 | 1.73E-02 | olanzapine 23 mg/kg | Group 5 |

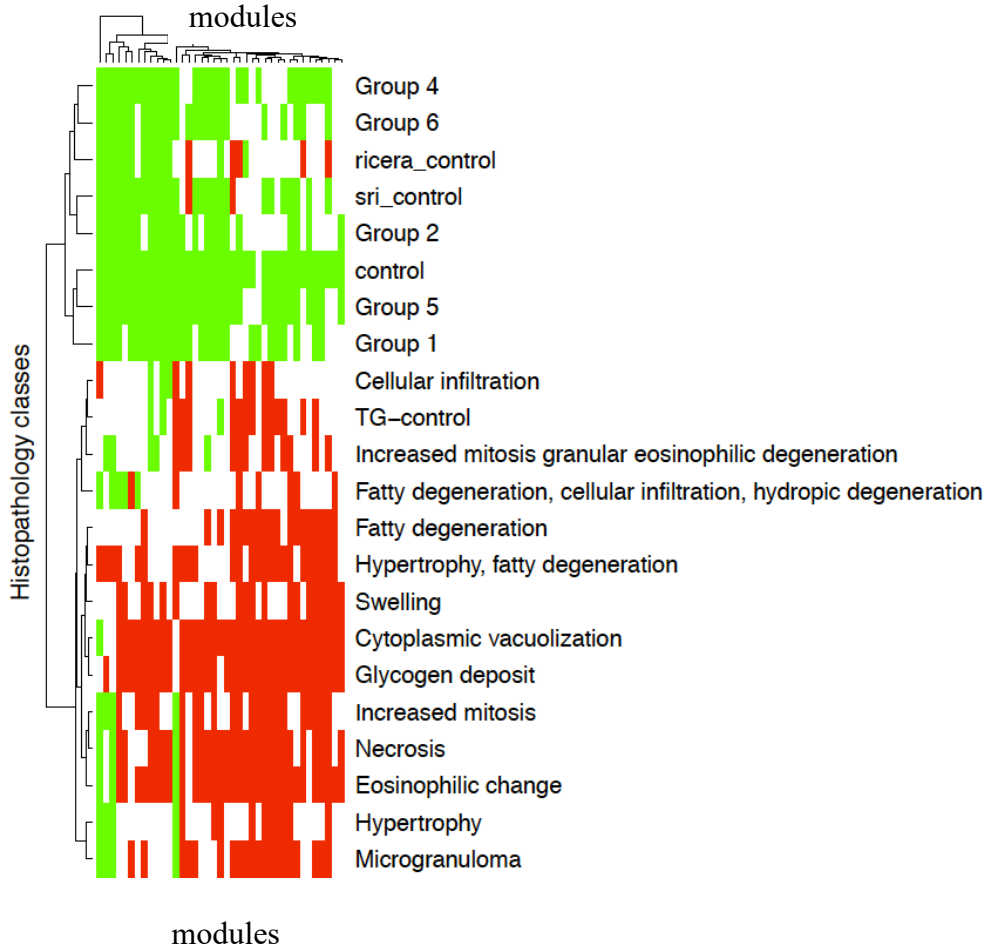
| | | | | | | |
|---|---|---|---|---|---|---|
| inclusion body assembly | | | | | | |
| GO:1901029 negative regulation of mitochondrial outer membrane permeabilization involved in apoptotic signaling pathway | 2 | 9 | 1.39E-05 | 1.73E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:1903265 positive regulation of tumor necrosis factor-mediated signaling pathway | 2 | 9 | 1.39E-05 | 1.73E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0010941 regulation of cell death | 2 | 11 | 2.12E-05 | 2.40E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0009612 response to mechanical stimulus | 3 | 101 | 3.20E-05 | 3.31E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0051131 chaperone-mediated protein complex assembly | 2 | 14 | 3.50E-05 | 3.35E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0033120 positive regulation of RNA splicing | 2 | 15 | 4.04E-05 | 3.35E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0006402 mRNA catabolic process | 2 | 15 | 4.04E-05 | 3.35E-02 | olanzapine 23 mg/kg | Group 5 |

| | | | | | | |
|---|---|---|---|---|---|---|
| GO:0031396 regulation of protein ubiquitination | 2 | 17 | 5.23E-05 | 3.42E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0034620 cellular response to unfolded protein | 2 | 17 | 5.23E-05 | 3.42E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:1902236 negative regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway | 2 | 17 | 5.23E-05 | 3.42E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0042026 protein refolding | 2 | 17 | 5.23E-05 | 3.42E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0035994 response to muscle stretch | 2 | 20 | 7.30E-05 | 4.32E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:1901673 regulation of mitotic spindle assembly | 2 | 20 | 7.30E-05 | 4.32E-02 | olanzapine 23 mg/kg | Group 5 |
| GO:0042744 hydrogen peroxide catabolic process | 4 | 21 | 1.15E-06 | 1.43E-02 | 17 methyltestosterone 2000 mg/kg | Group 5 |
| GO:0015671 oxygen transport | 3 | 10 | 6.50E-06 | 3.29E-02 | 17 methyltestosterone 2000 mg/kg | Group 5 |
| GO:0055114 oxidation-reduction process | 10 | 453 | 7.94E-06 | 3.29E-02 | 17 methyltestosterone 2000 mg/kg | Group 5 |
| GO:1902380 positive regulation of | 2 | 2 | 1.48E-05 | 4.60E-02 | 17 methyltestosterone 2000 mg/kg | Group 5 |

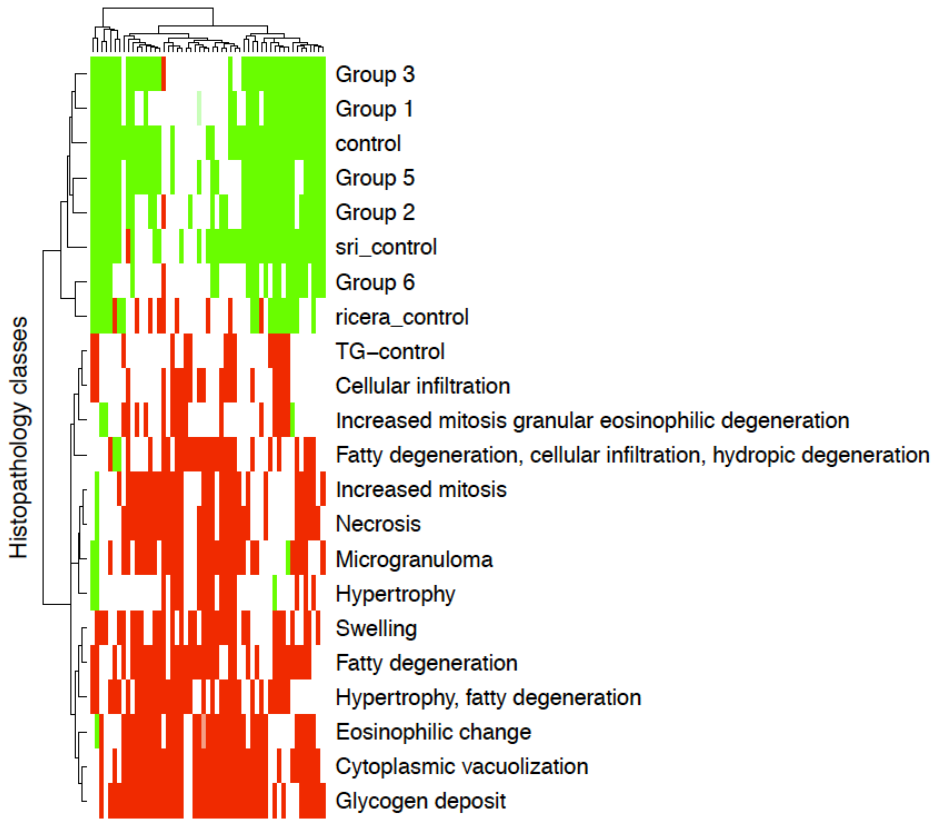
| | | | | | | |
|---|---|---|---|---|---|---|
| endoribonuclease activity | | | | | | |
| GO:0015671 oxygen transport | 2 | 10 | 2.90E-06 | 3.60E-02 | imatinib 150 mg/kg | Group 5 |
| PW:0000375 phase I biotransformation pathway via cytochrome P450 | 6 | 69 | 1.60E-06 | 1.60E-03 | carbamazepine 490 mg/kg | Group 5 |
| PW:0000054 tryptophan metabolic pathway | 4 | 38 | 4.57E-05 | 2.29E-02 | carbamazepine 490 mg/kg | Group 5 |
| PW:0000062 ascorbate and aldarate metabolic pathway | 3 | 16 | 7.48E-05 | 2.50E-02 | carbamazepine 490 mg/kg | Group 5 |
| PW:0000134 glutathione metabolic pathway | 4 | 47 | 1.07E-04 | 2.67E-02 | carbamazepine 490 mg/kg | Group 5 |
| GO:0055114 oxidation-reduction process | 13 | 454 | 6.03E-07 | 7.50E-03 | carbamazepine 490 mg/kg | Group 5 |
| PW:0000054 tryptophan metabolic pathway | 3 | 38 | 1.09E-05 | 1.09E-02 | bithionol 333 mg/kg | Group 6 |



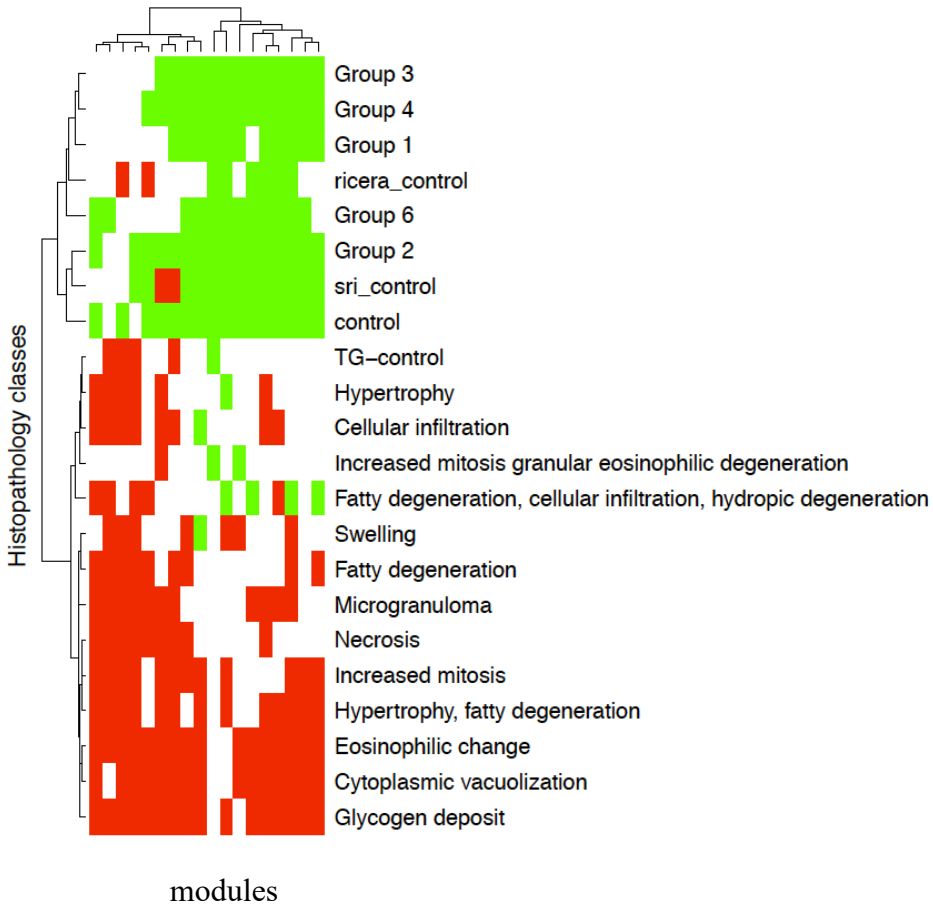

Group 2



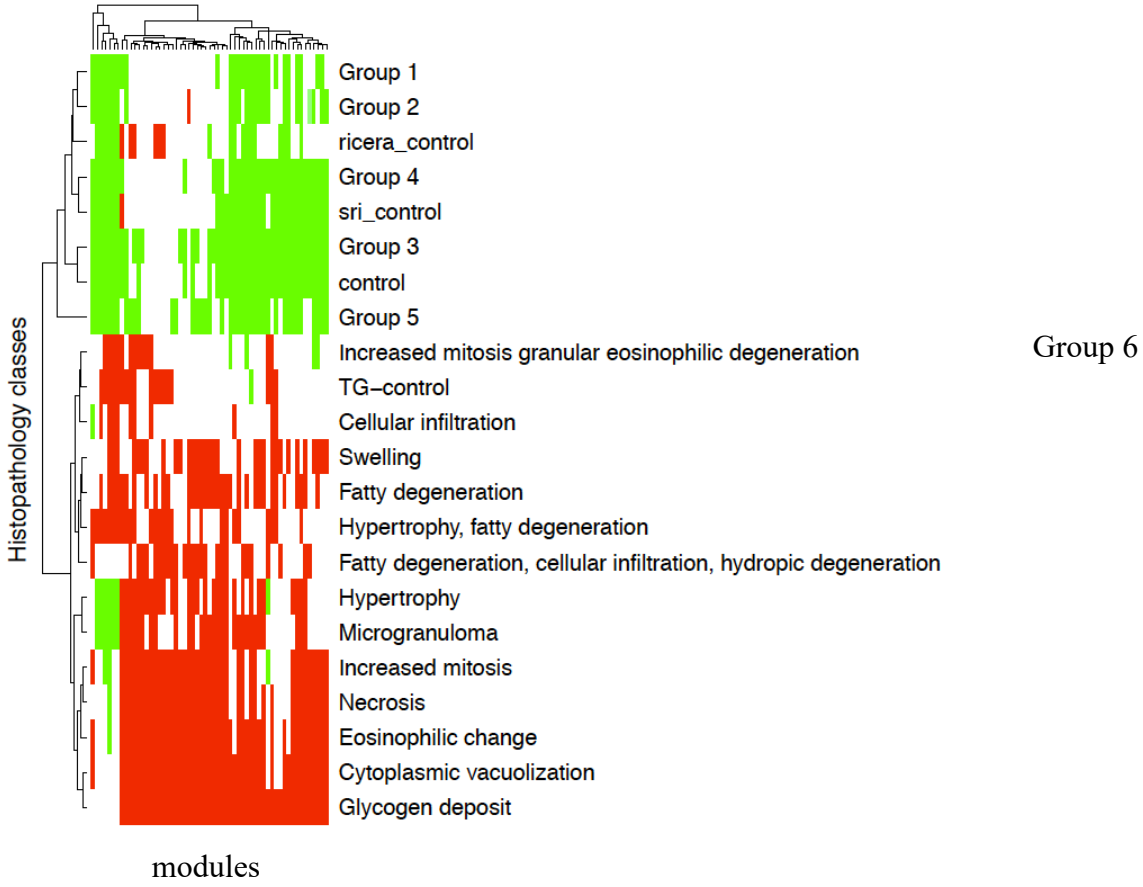

Group 3




Group 4




Group 5




*Supplementary Figure 2 shows the network conservation of modules for groups 2-6. In each case, DrugMatrix and Open TG-GATEs are separated. Modules which are conserved (green) are those of interest for pathway enrichment and biological signal deconvolution. These figures are analogous to Figure 3-5*

*Supplementary Table 4 Rat genome database (RGD) definition of the hub genes for Group 1's modules of interest. Hub genes were quantified using degree.*

| Module | name | Degree | RGD definition |
|---|---|---|---|
| darkgreen | Ftcd | 26 | ENCODES a protein that exhibits formimidoyltetrahydrofolate cyclodeaminase activity; formimidoyltransferase activity; microtubule binding; INVOLVED IN cytoskeleton organization; PARTICIPATES IN folate metabolic pathway; hereditary folate malabsorption pathway; histidine metabolic pathway; ASSOCIATED WITH Axenfeld-Rieger syndrome type 3 (ortholog); Bethlem myopathy (ortholog); Burkitt lymphoma (ortholog); FOUND IN endoplasmic reticulum; endoplasmic reticulum-Golgi intermediate compartment; Golgi apparatus; INTERACTS WITH 1,2-dimethylhydrazine; 1-benzylpiperazine; 1-naphthyl isothiocyanate |
| darkgreen | Ndufa5 | 20 | ENCODES a protein that exhibits NADH dehydrogenase (ubiquinone) activity (ortholog); INVOLVED IN mitochondrial respiratory chain complex I assembly (ortholog); respiratory electron transport chain (ortholog); PARTICIPATES IN Alzheimer's disease pathway; Huntington's disease pathway; oxidative phosphorylation pathway; ASSOCIATED WITH Facial Nerve Injuries; Alzheimer's disease (ortholog); pleomorphic xanthoastrocytoma (ortholog); FOUND IN mitochondrion; protein-containing complex; mitochondrial respiratory chain complex I (ortholog); INTERACTS WITH 17alpha-ethynylestradiol; 3,3',5-triiodo-L-thyronine; 6-propyl-2-thiouracil |
| darkgreen | Smarcd2 | 16 | INVOLVED IN chromatin remodeling (ortholog); nucleosome disassembly (ortholog); PARTICIPATES IN SWI/SNF family mediated chromatin remodeling pathway; ASSOCIATED WITH myeloid leukemia (ortholog); neutropenia (ortholog); Smith-Magenis syndrome (ortholog); FOUND IN nucleoplasm (ortholog); SWI/SNF complex (ortholog); INTERACTS WITH (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 17alpha-ethynylestradiol; 17beta-estradiol |
| darkgreen | Ube2k | 16 | ENCODES a protein that exhibits ubiquitin protein ligase binding (ortholog); ubiquitin-protein transferase activity (ortholog); ubiquitin-ubiquitin ligase activity (ortholog); INVOLVED IN intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress; cellular response to interferon-beta (ortholog); free ubiquitin chain polymerization (ortholog); PARTICIPATES IN ubiquitin/proteasome degradation pathway; FOUND IN cytoplasm (ortholog); filopodium tip (ortholog); nucleus (ortholog); INTERACTS WITH 2,4-dinitrotoluene; 2,6-dinitrotoluene; bisphenol A |
| darkgreen | Ass1 | 11 | ENCODES a protein that exhibits argininosuccinate synthase activity; toxic substance binding; amino acid binding (ortholog); INVOLVED IN acute-phase response; aging; arginine biosynthetic process; PARTICIPATES IN urea cycle pathway; AGAT deficiency pathway; arginine and proline metabolic pathway; ASSOCIATED WITH Acute-Phase Reaction; Diabetes Mellitus, Experimental ; Drug Toxicity; FOUND IN cell body fiber; cytoplasm; endoplasmic reticulum; INTERACTS WITH (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 17alpha-ethynylestradiol; 2,2,2-tetramine |
| cyan | Cxcl2 | 12 | ENCODES a protein that exhibits chemokine activity; INVOLVED IN cellular response to interleukin-1; cellular response to lipopolysaccharide; leukocyte chemotaxis; PARTICIPATES IN chemokine mediated signaling pathway; cytokine mediated signaling pathway; NOD-like receptor signaling pathway; ASSOCIATED WITH Acute Lung Injury; Acute Necrotizing Pancreatitis; Alcoholic Liver Diseases; FOUND IN extracellular space; INTERACTS WITH (+)-alpha-tocopherol; (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 1,2,4-trimethylbenzene |

| | | | |
|---|---|---|---|
| cyan | Il10 | 11 | ENCODES a protein that exhibits interleukin-10 receptor binding; cytokine activity (ortholog); INVOLVED IN aging; cellular response to estradiol stimulus; immune response; PARTICIPATES IN Interleukin-10 signaling pathway; interleukin-4 signaling pathway; allograft rejection pathway; ASSOCIATED WITH Acute Lung Injury; adult respiratory distress syndrome; Alveolar Bone Loss; FOUND IN extracellular space; INTERACTS WITH (+)-alpha-tocopherol; (+)-pilocarpine; (-)-epigallocatechin 3-gallate |
| cyan | Nr4a2 | 8 | ENCODES a protein that exhibits proximal promoter DNA-binding transcription activator activity, RNA polymerase II-specific; RNA polymerase II proximal promoter sequence-specific DNA binding; sequence-specific DNA binding; INVOLVED IN positive regulation of transcription by RNA polymerase II; response to inorganic substance; response to insecticide; PARTICIPATES IN Parkinson's disease pathway; ASSOCIATED WITH Arsenic Poisoning (ortholog); autism spectrum disorder (ortholog); autistic disorder (ortholog); FOUND IN cytoplasm; nucleus; nuclear speck (ortholog); INTERACTS WITH 1,2,4-trimethylbenzene; 3,7-dihydropurine-6-thione; 6-propyl-2-thiouracil |
| cyan | Dusp2 | 5 | ENCODES a protein that exhibits mitogen-activated protein kinase binding (ortholog); phosphoprotein phosphatase activity (ortholog); INVOLVED IN protein dephosphorylation (ortholog); PARTICIPATES IN mitogen activated protein kinase signaling pathway; ASSOCIATED WITH juvenile rheumatoid arthritis (ortholog); schizophrenia (ortholog); FOUND IN nuclear membrane (ortholog); nucleus (ortholog); INTERACTS WITH (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 6-propyl-2-thiouracil; acrylamide |
| cyan | Il1r2 | 5 | ENCODES a protein that exhibits interleukin-1 binding (ortholog); interleukin-1 receptor activity (ortholog); INVOLVED IN negative regulation of cytokine production involved in inflammatory response (ortholog); negative regulation of interleukin-1 alpha secretion (ortholog); negative regulation of interleukin-1-mediated signaling pathway (ortholog); PARTICIPATES IN interleukin-1 signaling pathway; cytokine mediated signaling pathway; Entamoebiasis pathway; ASSOCIATED WITH aggressive periodontitis (ortholog); allergic hypersensitivity disease (ortholog); ankylosing spondylitis (ortholog); FOUND IN cytoplasm (ortholog); INTERACTS WITH (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 1-benzylpiperazine; 2-amino-2-deoxy-D-glucopyranose |
| lightgreen | Pcyt1a | 52 | ENCODES a protein that exhibits calmodulin binding; choline-phosphate cytidylyltransferase activity; lipid binding; INVOLVED IN CDP-choline pathway; phosphatidylcholine biosynthetic process; PARTICIPATES IN glycerophospholipid metabolic pathway; lamivudine pharmacokinetics pathway; ASSOCIATED WITH Chemical and Drug Induced Liver Injury (ortholog); genetic disease (ortholog); schizophrenia (ortholog); FOUND IN cytoplasm; endoplasmic reticulum; nuclear envelope; INTERACTS WITH (R)-mevalonic acid; (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 2,4-dinitrotoluene |
| lightgreen | Map1lc3b | 51 | ENCODES a protein that exhibits microtubule binding; protein domain specific binding; tubulin binding; INVOLVED IN positive regulation of protein binding; autophagosome maturation (ortholog); autophagy (ortholog); PARTICIPATES IN autophagy pathway; ASSOCIATED WITH Myocardial Ischemia; 16Q24.3 Microdeletion Syndrome (ortholog); Chemical and Drug Induced Liver Injury (ortholog); FOUND IN autophagosome membrane; axon; cytoplasm; INTERACTS WITH 2,2',5,5'-tetrachlorobiphenyl; 2,3,7,8-tetrachlorodibenzodioxine; 3-methyladenine |
| lightgreen | Rhbdd2 | 51 | ENCODES a protein that exhibits serine-type endopeptidase activity (inferred); INVOLVED IN proteolysis (inferred); ASSOCIATED WITH pleomorphic xanthoastrocytoma (ortholog); FOUND IN Golgi apparatus (ortholog); Golgi membrane (ortholog); nucleoplasm (ortholog); INTERACTS WITH 11-CYCLOPROPYL-5,11-DIHYDRO-4-METHYL-6H-DIPYRIDO[3,2-B:2',3'-E][1,4]DIAZEPIN-6-ONE; 2,3,7,8-tetrachlorodibenzodioxine; bisphenol A |

| | | | |
|---|---|---|---|
| lightgreen | Fkbp5 | 49 | ENCODES a protein that exhibits heat shock protein binding (ortholog); peptidyl-prolyl cis-trans isomerase activity (ortholog); INVOLVED IN chaperone-mediated protein folding (ortholog); response to bacterium (ortholog); PARTICIPATES IN aldosterone signaling pathway; cortisol signaling pathway; ASSOCIATED WITH endogenous depression (ortholog); endometriosis (ortholog); Infant, Newborn, Diseases (ortholog); FOUND IN nucleoplasm (ortholog); INTERACTS WITH (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 11-CYCLOPROPYL-5,11-DIHYDRO-4-METHYL-6H-DIPYRIDO[3,2-B:2',3'-E][1,4]DIAZEPIN-6-ONE; 2,3,7,8-tetrachlorodibenzodioxine |
| lightgreen | Tmem38b | 44 | ENCODES a protein that exhibits potassium channel activity (inferred); INVOLVED IN potassium ion transmembrane transport (inferred); ASSOCIATED WITH Chemical and Drug Induced Liver Injury (ortholog); osteogenesis imperfecta type 14 (ortholog); schizophrenia (ortholog); FOUND IN endoplasmic reticulum membrane (inferred); integral component of membrane (inferred); INTERACTS WITH 1,3-dinitrobenzene; 2,3,7,8-tetrachlorodibenzodioxine; acrylamide |
| Brown | Orc4 | 15 | ENCODES a protein that exhibits DNA replication origin binding (ortholog); nucleotide binding (ortholog); INVOLVED IN DNA replication initiation (ortholog); ASSOCIATED WITH autosomal dominant non-syndromic intellectual disability 1 (ortholog); Meier-Gorlin syndrome (ortholog); schizophrenia (ortholog); FOUND IN cytosol (ortholog); nuclear chromosome, telomeric region (ortholog); nuclear origin of replication recognition complex (ortholog); INTERACTS WITH 2,4-dinitrotoluene; 2,6-dinitrotoluene; bis(2-ethylhexyl) phthalate |
| Brown | Atp6v1h | 11 | ENCODES a protein that exhibits proton-transporting ATPase activity, rotational mechanism (inferred); INVOLVED IN endocytosis (ortholog); PARTICIPATES IN oxidative phosphorylation pathway; phagocytosis pathway; rheumatoid arthritis pathway; FOUND IN vacuolar proton-transporting V-type ATPase, V1 domain (inferred); INTERACTS WITH bisphenol A; cadmium dichloride; flutamide |
| Brown | Tbc1d15 | 10 | ENCODES a protein that exhibits GTPase activator activity (ortholog); INVOLVED IN regulation of GTPase activity (ortholog); FOUND IN cytoplasm (ortholog); extracellular region (ortholog); INTERACTS WITH bisphenol A; cobalt dichloride; flutamide |
| Brown | Slc25a17 | 9 | ENCODES a protein that exhibits ADP transmembrane transporter activity (ortholog); AMP transmembrane transporter activity (ortholog); ATP transmembrane transporter activity (ortholog); INVOLVED IN ATP transport (ortholog); fatty acid beta-oxidation (ortholog); fatty acid transport (ortholog); FOUND IN integral component of peroxisomal membrane (ortholog); peroxisomal membrane (ortholog); peroxisome (ortholog); INTERACTS WITH 2,6-dinitrotoluene; bisphenol A; buspirone |
| Brown | Lamtor3 | 8 | ENCODES a protein that exhibits guanyl-nucleotide exchange factor activity (ortholog); kinase activator activity (ortholog); protein-containing complex scaffold activity (ortholog); INVOLVED IN activation of MAPKK activity (ortholog); cellular protein localization (ortholog); cellular response to amino acid stimulus (ortholog); PARTICIPATES IN mTOR signaling pathway; FOUND IN late endosome (ortholog); Ragulator complex (ortholog); INTERACTS WITH 2,6-dinitrotoluene; dibutyl phthalate; flutamide |
| magenta | Rpl17 | 51 | ENCODES a protein that exhibits large ribosomal subunit rRNA binding; INVOLVED IN cellular response to amino acid starvation; positive regulation of G1/S transition of mitotic cell cycle; response to amino acid starvation; PARTICIPATES IN ribosome biogenesis pathway; translation pathway; ASSOCIATED WITH Myocardial Ischemia (ortholog); FOUND IN A band; cytosolic large ribosomal subunit; nucleus; INTERACTS WITH 2,4-dinitrotoluene; 2,6-dinitrotoluene; 2-nitrofluorene |
| magenta | Rpl15 | 50 | INVOLVED IN response to ethanol; PARTICIPATES IN ribosome biogenesis pathway; translation pathway; ASSOCIATED WITH Diamond-Blackfan anemia (ortholog); Diamond-Blackfan Anemia 12 (ortholog); Disease Progression (ortholog); FOUND IN A band; cytosolic large ribosomal subunit; nucleus (ortholog); INTERACTS WITH 17alpha-ethynylestradiol; 2,4-dinitrotoluene; 2,6-dinitrotoluene |

| | | | |
|---|---|---|---|
| magenta | Rpl5 | 46 | ENCODES a protein that exhibits 5S rRNA binding; mRNA binding; mRNA 3'-UTR binding (ortholog); INVOLVED IN cellular response to inorganic substance; positive regulation of isoleucine-tRNA ligase activity; positive regulation of methionine-tRNA ligase activity; PARTICIPATES IN ribosome biogenesis pathway; translation pathway; ASSOCIATED WITH aplastic anemia (ortholog); Diamond-Blackfan anemia (ortholog); Diamond-Blackfan Anemia 6 (ortholog); FOUND IN aminoacyl-tRNA synthetase multienzyme complex; cytosolic large ribosomal subunit; cytosolic ribosome; INTERACTS WITH 17alpha-ethynylestradiol; 2,3,7,8-tetrachlorodibenzodioxine; ammonium chloride |
| magenta | Rps12 | 44 | ENCODES a protein that exhibits structural constituent of ribosome (inferred); INVOLVED IN response to organonitrogen compound; PARTICIPATES IN ribosome biogenesis pathway; translation pathway; ASSOCIATED WITH Hypertriglyceridemia; FOUND IN cytosolic large ribosomal subunit; cytosolic small ribosomal subunit; cytosol (ortholog); INTERACTS WITH 2,6-dinitrotoluene; 3H-1,2-dithiole-3-thione; ammonium chloride |
| magenta | Rps8 | 39 | INVOLVED IN translation (inferred); PARTICIPATES IN ribosome biogenesis pathway; translation pathway; ASSOCIATED WITH Breast Neoplasms (ortholog); Charcot-Marie-Tooth disease dominant intermediate C (ortholog); Parkinson's disease (ortholog); FOUND IN cytosolic small ribosomal subunit; cytosol (ortholog); endoplasmic reticulum (ortholog); INTERACTS WITH 1,1,1-Trichloro-2-(o-chlorophenyl)-2-(p-chlorophenyl)ethane; 1-naphthyl isothiocyanate; 17alpha-ethynylestradiol |
| yellow | Cdk1 | 56 | ENCODES a protein that exhibits cyclin binding; cyclin-dependent protein serine/threonine kinase activity; histone kinase activity; INVOLVED IN animal organ regeneration; cell aging; cellular response to hydrogen peroxide; PARTICIPATES IN cell cycle pathway, mitotic; G1/S transition pathway; G2/M checkpoint pathway; ASSOCIATED WITH hepatocellular carcinoma; Spinal Cord Injuries; Thyroid Neoplasms; FOUND IN cytoplasm; nucleus; centrosome (ortholog); INTERACTS WITH (+)-catechin; (+)-pilocarpine; (R)-mevalonic acid |
| yellow | Nusap1 | 56 | ENCODES a protein that exhibits microtubule binding (ortholog); INVOLVED IN establishment of mitotic spindle localization (ortholog); mitotic chromosome condensation (ortholog); mitotic cytokinesis (ortholog); ASSOCIATED WITH Colorectal Neoplasms (ortholog); FOUND IN nucleolus (ortholog); spindle microtubule (ortholog); INTERACTS WITH (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 2,3,7,8-tetrachlorodibenzodioxine; 2-acetamidofluorene |
| yellow | Spc25 | 53 | INVOLVED IN chromosome segregation (ortholog); mitotic spindle organization (ortholog); FOUND IN condensed chromosome kinetochore (ortholog); cytosol (ortholog); Ndc80 complex (ortholog); INTERACTS WITH (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 17alpha-ethynylestradiol; 2,3,7,8-tetrachlorodibenzodioxine |
| yellow | Ube2c | 52 | ENCODES a protein that exhibits ubiquitin conjugating enzyme activity (ortholog); ubiquitin-like protein ligase binding (ortholog); ubiquitin-protein transferase activity (ortholog); INVOLVED IN anaphase-promoting complex-dependent catabolic process (ortholog); exit from mitosis (ortholog); free ubiquitin chain polymerization (ortholog); PARTICIPATES IN ubiquitin/proteasome degradation pathway; ASSOCIATED WITH Breast Neoplasms (ortholog); Chromosome Aberrations (ortholog); FOUND IN anaphase-promoting complex (ortholog); cytosol (ortholog); plasma membrane (ortholog); INTERACTS WITH (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 1,3-dinitrobenzene; 2,3,7,8-tetrachlorodibenzodioxine |
| yellow | Ccnb1 | 49 | ENCODES a protein that exhibits histone kinase activity; protein kinase binding; protein-containing complex binding; INVOLVED IN cellular response to fatty acid; cellular response to hypoxia; cellular response to iron(III) ion; PARTICIPATES IN cell cycle pathway, mitotic; G1/S transition pathway; G2/M checkpoint pathway; ASSOCIATED WITH Diabetes Mellitus, Experimental ; |

| | | | |
|---|---|---|---|
| | | | Experimental Liver Neoplasms; Hyperplasia; FOUND IN cytoplasm; nucleus; centrosome (ortholog); INTERACTS WITH (-)-citrinin; (S)-10-[(DIMETHYLAMINO)METHYL]-4-ETHYL-4,9-DIHYDROXY-1H-PYRANO[3',4':6,7]INOLIZINO[1,2-B]-QUINOLINE-3,14(4H,12H)-DIONE; 1,1,1-Trichloro-2-(o-chlorophenyl)-2-(p-chlorophenyl)ethane |
| turquoise | Relt | 132 | INVOLVED IN apoptotic process (ortholog); PARTICIPATES IN cytokine mediated signaling pathway; FOUND IN nucleus (ortholog); INTERACTS WITH 6-propyl-2-thiouracil; acrylamide; bisphenol A |
| turquoise | RGD1312005 | 109 | #N/A |
| turquoise | Cnksr1 | 95 | ENCODES a protein that exhibits protein binding, bridging (ortholog); INVOLVED IN Ras protein signal transduction (ortholog); Rho protein signal transduction (ortholog); PARTICIPATES IN angiotensin II signaling pathway via AT2 receptor; interleukin-3 signaling pathway; ASSOCIATED WITH intellectual disability (ortholog); FOUND IN cell cortex (ortholog); INTERACTS WITH 2,3,7,8-tetrachlorodibenzodioxine; 3-chloropropane-1,2-diol; 6-propyl-2-thiouracil |
| turquoise | Epb41l4a | 51 | ASSOCIATED WITH Failure to Thrive (ortholog); Hereditary Neoplastic Syndromes (ortholog); FOUND IN cytoskeleton (inferred); INTERACTS WITH aflatoxin B1; all-trans-retinoic acid; bisphenol A |
| turquoise | LOC56764 | 49 | #N/A |

Supplementary Table 5:Enriched pathway and GO terms for each transformation and their overlap. All gene sets are significantly enriched with an FDR-adjusted p value < 0.05.

| Pathways | GO terms |
|---|---|
| G1_1day->G1 | |
| PW:0000015 Alzheimer's disease pathway | GO:0000082 G1/S transition of mitotic cell cycle |
| PW:0000017 Huntington's disease pathway | GO:0006099 tricarboxylic acid cycle |
| PW:0000018 Parkinson's disease pathway | GO:0006103 2-oxoglutarate metabolic process |
| PW:0000025 glycolysis/gluconeogenesis pathway | GO:0006260 DNA replication |
| PW:0000026 citric acid cycle pathway | GO:0006261 DNA-dependent DNA replication |
| PW:0000034 electron transport chain pathway | GO:0006270 DNA replication initiation |
| PW:0000045 pentose phosphate pathway | GO:0009060 aerobic respiration |
| PW:0000398 homocysteine metabolic pathway | GO:0009615 response to virus |
| PW:0000640 glycolysis pathway | GO:0009617 response to bacterium |
| PW:0000641 gluconeogenesis pathway | GO:0022900 electron transport chain |
| PW:0000662 mismatch repair pathway | GO:0032981 mitochondrial respiratory chain complex I assembly |
| PW:0000718 p53 signaling pathway | GO:0035458 cellular response to interferon-beta |
| PW:0000817 NOD-like receptor signaling pathway | GO:0045071 negative regulation of viral genome replication |
| PW:0001059 oxidative phosphorylation pathway | GO:0051321 meiotic cell cycle |
| PW:0001078 cysteine and methionine metabolic pathway | GO:0051607 defense response to virus |
| PW:0001610 tyrosinemia type III pathway | GO:2000059 negative regulation of ubiquitin-dependent protein catabolic process |
| PW:0001628 triosephosphate isomerase deficiency pathway | |
| PW:0001754 pyruvate dehydrogenase E2 deficiency pathway | |
| PW:0001755 pyruvate dehydrogenase E3 deficiency pathway | |
| PW:0001814 mitochondrial complex II deficiency pathway | |
| PW:0001992 glycogen storage disease type Ia pathway | |
| PW:0001993 glycogen storage disease type Ib pathway | |
| PW:0002098 fumaric aciduria pathway | |
| PW:0002100 fructose-1,6-bisphosphatase deficiency pathway | |
| PW:0002617 phosphoenolpyruvate carboxykinase deficiency pathway | |
| | |
| G2_1day->G2 | |
| PW:0000180 mTOR signaling pathway | GO:0006955 immune response |

| | |
|---|---|
| | |
| G3_1day->G3 | |
| PW:0000019 prion disease pathway | GO:0001732 formation of cytoplasmic translation initiation complex |
| PW:0000126 RNA polymerase I transcription pathway | GO:0006749 glutathione metabolic process |
| PW:0000134 glutathione metabolic pathway | GO:0006954 inflammatory response |
| PW:0000829 chemokine mediated signaling pathway | GO:0007166 cell surface receptor signaling pathway |
| PW:0001023 systemic lupus erythematosus pathway | GO:0030593 neutrophil chemotaxis |
| PW:0001146 Fc gamma receptor mediated signaling pathway | GO:0042102 positive regulation of T cell proliferation |
| | GO:0042176 regulation of protein catabolic process |
| | GO:0042254 ribosome biogenesis |
| | GO:0042535 positive regulation of tumor necrosis factor biosynthetic process |
| | GO:0045454 cell redox homeostasis |
| | GO:0045727 positive regulation of translation |
| | GO:0045899 positive regulation of RNA polymerase II transcriptional preinitiation complex assembly |
| | GO:0050790 regulation of catalytic activity |
| | GO:0071222 cellular response to lipopolysaccharide |
| | GO:0071346 cellular response to interferon-gamma |
| | GO:1901800 positive regulation of proteasomal protein catabolic process |
| | GO:2000249 regulation of actin cytoskeleton reorganization |
| | GO:2000406 positive regulation of T cell migration |
| | |
| G4_1day->G4 | |
| PW:0000127 RNA polymerase II transcription pathway | GO:0000387 spliceosomal snRNP assembly |
| PW:0000129 base excision repair pathway | GO:0006606 protein import into nucleus |
| PW:0000202 homologous recombination pathway of double-strand break repair | GO:0006913 nucleocytoplasmic transport |
| PW:0001151 Fc epsilon receptor mediated signaling pathway | |
| PW:0001587 mRNA nuclear export pathway | |
| PW:0001626 CRM1 export pathway | |

| | |
|---|---|
| G5_1day->G5 | |
| PW:0000050 arginine and proline metabolic pathway | GO:0006468 protein phosphorylation |
| PW:0000088 G1/S transition pathway | GO:0007019 microtubule depolymerization |
| PW:0000214 polyamine metabolic pathway | GO:0007080 mitotic metaphase plate congression |
| PW:0000373 glutathione conjugation pathway | GO:0007088 regulation of mitotic nuclear division |
| PW:0000528 angiotensin II signaling pathway via AT2 receptor | GO:0032465 regulation of cytokinesis |
| PW:0000638 Endoplasmic Reticulum-associated degradation pathway | GO:0032467 positive regulation of cytokinesis |
| | |
| G6_1day->G6 | |
| PW:0000008 Wnt signaling pathway | GO:0006360 transcription by RNA polymerase I |
| PW:0000862 de novo pyrimidine biosynthetic pathway | GO:0007030 Golgi organization |
| PW:0001406 SWI/SNF family mediated chromatin remodeling pathway | |
| PW:0001944 mercaptopurine pharmacodynamics pathway | |
| PW:0002203 azathioprine pharmacodynamics pathway | |
| PW:0002394 tioguanine pharmacodynamics pathway | |
| | |
| G1_1day->G1 + G3_1day->G3 | |
| PW:0001159 aminoacyl-tRNA biosynthetic pathway | GO:0097421 liver regeneration |
| | |
| G1_1day->G1 + G4_1day->G4 | |
| PW:0000098 DNA replication pathway | GO:0071897 DNA biosynthetic process |
| PW:0000130 nucleotide excision repair pathway | |
| | |
| G1_1day->G1 + G5_1day->G5 | |
| PW:0000086 cell cycle pathway, mitotic | GO:0000070 mitotic sister chromatid segregation |
| PW:0000759 gemcitabine pharmacokinetics pathway | GO:0000278 mitotic cell cycle |
| PW:0000760 gemcitabine pharmacodynamics pathway | GO:0000281 mitotic cytokinesis |
| | GO:0006281 DNA repair |
| | GO:0007018 microtubule-based movement |
| | GO:0007051 spindle organization |
| | GO:0007052 mitotic spindle organization |
| | GO:0007059 chromosome segregation |
| | GO:0007094 mitotic spindle assembly checkpoint |

| | |
|---|---|
| | GO:0008283 cell proliferation |
| | GO:0051301 cell division |
| | |
| G1_1day->G1 + G6_1day->G6 | |
| PW:0000048 methionine cycle/metabolic pathway | GO:0000462 maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) |
| PW:0001073 spliceosome pathway | |
| | |
| G3_1day->G3 + G5_1day->G5 | |
| PW:0000144 ubiquitin/proteasome degradation pathway | GO:0010498 proteasomal protein catabolic process |
| PW:0000375 phase I biotransformation pathway via cytochrome P450 | GO:0010499 proteasomal ubiquitin-independent protein catabolic process |
| PW:0000474 coagulation cascade pathway | GO:0043161 proteasome-mediated ubiquitin-dependent protein catabolic process |
| PW:0000502 complement system pathway | |
| | |
| G3_1day->G3 + G6_1day->G6 | |
| | GO:0006413 translational initiation |
| | |
| G4_1day->G4 + G6_1day->G6 | |
| PW:0000031 purine metabolic pathway | |
| PW:0000867 de novo purine biosynthetic pathway | |
| PW:0001590 xanthinuria pathway | |
| PW:0001591 xanthinuria type I pathway | |
| PW:0001592 xanthinuria type II pathway | |
| PW:0001777 purine nucleoside phosphorylase deficiency pathway | |
| PW:0001779 adenosine monophosphate deaminase deficiency pathway | |
| PW:0001817 molybdenum cofactor deficiency pathway | |
| PW:0001879 Lesch-Nyhan syndrome pathway | |
| PW:0001938 Kelley-Seegmiller syndrome pathway | |
| PW:0002137 presequence pathway of mitochondrial protein import | |
| PW:0002140 beta-barrel pathway of mitochondrial protein import | |
| PW:0002276 adenylosuccinate lyase deficiency pathway | |
| PW:0002294 AICA-ribosuria pathway | |
| PW:0002566 adenine phoshoribosyltransferase deficiency pathway | |
| | |
| G1_1day->G1 + G3_1day->G3 + G5_1day->G5 | |
| | GO:0000028 ribosomal small subunit assembly |
| | GO:0002181 cytoplasmic translation |
| | GO:0042274 ribosomal small subunit biogenesis |

| | |
|---|---|
| | GO:0045087 innate immune response |
| | |
| | |
| G1_1day->G1 + G3_1day->G3 + G6_1day->G6 | |
| PW:0000580 translation initiation pathway | GO:0000463 maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) |
| | |
| G2_1day->G2 + G3_1day->G3 + G5_1day->G5 | |
| PW:0001045 Staphylococcus aureus infection pathway | |
| | |
| G1_1day->G1 + G3_1day->G3 + G4_1day->G4 + G6_1day->G6 | |
| | GO:0008150 biological_process |
| | |
| G1_1day->G1 + G3_1day->G3 + G5_1day->G5 + G6_1day->G6 | |
| | GO:0000027 ribosomal large subunit assembly |
| | GO:0006364 rRNA processing |
| | |
| G1_1day->G1 + G4_1day->G4 + G5_1day->G5 + G6_1day->G6 | |
| PW:0000032 pyrimidine metabolic pathway | |
| | |
| G1_1day->G1 + G3_1day->G3 + G4_1day->G4 + G5_1day->G5 + G6_1day->G6 | |
| PW:0000101 translation pathway | GO:0006412 translation |
| PW:0001066 ribosome biogenesis pathway | GO:0042273 ribosomal large subunit biogenesis |
| | |
| G1_1day->G1 + G2_1day->G2 + G3_1day->G3 + G4_1day->G4 + G5_1day->G5 + G6_1day->G6 | |
| PW:0001160 RNA transport pathway | |